



Behavioral signatures of the rapid recruitment of long-term memory to overcome working memory capacity limits

Kirsten C. S. Adam¹ · Chong Zhao² · Edward K. Vogel²

Accepted: 28 March 2024
© The Psychonomic Society, Inc. 2024

Abstract

Working- and long-term memory are often studied in isolation. To better understand the specific limitations of working memory, effort is made to reduce the potential influence of long-term memory on performance in working memory tasks (e.g., asking participants to remember artificial, abstract items rather than familiar real-world objects). However, in everyday life we use working- and long-term memory in tandem. Here, our goal was to characterize how long-term memory can be recruited to circumvent capacity limits in a typical visual working memory task (i.e., remembering colored squares). Prior work has shown that incidental repetitions of working memory arrays often do not improve visual working memory performance – even after dozens of incidental repetitions, working memory performance often shows no improvement for repeated arrays. Here, we used a whole-report working memory task with explicit rather than incidental repetitions of arrays. In contrast to prior work with incidental repetitions, in two behavioral experiments we found that explicit repetitions of arrays yielded robust improvement to working memory performance, even after a single repetition. Participants performed above chance at recognizing repeated arrays in a later long-term memory test, consistent with the idea that long-term memory was used to rapidly improve performance across array repetitions. Finally, we analyzed inter-item response times and we found a response time signature of chunk formation that only emerged after the array was repeated (inter-response time slowing after two to three items); thus, inter-item response times may be useful for examining the coordinated interaction of visual working and long-term memory in future work.

Keywords Working memory · Long-term memory · Learning · Inter-response times

Introduction

Because of working memory's limited capacity, we need to flexibly recruit long-term memory to accomplish our everyday goals. For example, when performing a routine grocery shopping trip, it is not feasible to hold 20 different items actively in working memory. However, by taking advantage of associations in long-term memory, we can strategically retrieve “chunks” of items to effectively shop for all 20 items (e.g., “I should buy the ingredients I need to make caesar salad, lasagna, and tiramisù”) (Bower, 1972; Cowan, 2001; Ebbinghaus, 1885, 1913). Arguably, scenarios

like this one are the most common way that we use working memory in the real world. We frequently need to flexibly shuttle information back and forth between working- and long-term memory. Doing so allows us to take advantage of the strengths of each memory system — Information in working memory is easily accessible and manipulable but capacity-limited (Baddeley & Hitch, 1974; Cowan, 2001; Luck & Vogel, 1997), whereas information in long-term memory is nearly capacity unlimited but often takes time and effort to retrieve (Beck & van Lamsweerde, 2011; Brady et al., 2008; Mandler & Ritchey, 1977; Squire & Zola, 1996; Standing, 1973; Standing et al., 1970; Wolfe et al., 2023).

✉ Kirsten C. S. Adam
kadam@rice.edu

¹ Department of Psychological Sciences, Rice University,
6100 Main St, Houston, TX MS-651, USA

² Department of Psychology, University of Chicago, Chicago,
IL, USA

Probing interactions of visual working and long-term memory

Although working and long-term memory are typically used in tandem, researchers make an effort to devise working memory experiments that prevent the contribution of

long-term memory so that the unique constraints of working memory can be better characterized. For example, visual working memory is often studied by asking people to remember simple shapes or colors. Because these abstract items have low meaningfulness, these tasks help us to estimate the capacity of working memory in the absence of support from long-term memory associations.¹ Over the past many decades, careful work dissociating working and long-term memory has been important for our understanding of these memory systems and their neural correlates (Baddeley & Warrington, 1970; Christophel et al., 2017; Jeneson & Squire, 2011; Luck & Vogel, 1997; Milner & Penfield, 1955; Scoville & Milner, 1957; Serences et al., 2009). However, by focusing primarily on each memory system in isolation, we may miss important insights about how information is flexibly shuttled between both working and long-term memory in everyday life. One approach to closing this gap has been to use more realistic stimuli in working memory tasks, such as photographs of real-world objects and scenes (Brady et al., 2016; Brady & Störmer, 2022; Endress & Potter, 2014; Quirk et al., 2020). One advantage of these stimuli is that they may allow participants to draw on long-term memory via familiarity and meaningful associations (Asp et al., 2021; Jackson & Raymond, 2008; Ngiam et al., 2019b; Reder et al., 2013; Xie & Zhang, 2017, 2022). However, a disadvantage of using real-world objects is that they have associations in long-term memory that are already pre-formed, and the experimenter cannot directly control or observe how the formation of long-term memory associations influences working memory performance.

Rather than changing the memoranda to be more realistic, a second approach for studying interactions of working and long-term memory is to use artificial stimuli, but to introduce controlled opportunities for long-term memory to aid performance. In this vein, prior work has examined how incidental repetitions of memoranda may improve visual working memory performance (i.e., via Hebbian or implicit learning). Surprisingly, initial work on this topic found that visual working memory capacity was stubbornly resistant to improvement (Fukuda & Vogel, 2019; Logie et al., 2009; Olson & Jiang, 2004). For example, Olson and Jiang (2004) found that even after 24 repetitions of the same memory array, participants performed no better than as if they were

seeing the array for the first time. The lack of effect of repetitions on visual working memory performance is puzzling, because it contrasts with a rich body of work that shows that memory for verbal memoranda is improved with incidental repetitions (Hebb, 1961; Page et al., 2013; Sukegawa et al., 2019). As such, recent work has begun to systematically explore which factors may prevent versus allow Hebbian learning from incidental repetitions of visual working memory arrays (Musfeld et al., 2023a, 2023b; Souza & Oberauer, 2022). For example, Musfeld et al. (2023b) found that retrieval practice and the expected difficulty of the test both influence whether or not working memory performance improves when arrays are incidentally repeated over time.

Here, we turned our focus from incidental to explicit repetitions of working memory arrays. Explicit repetition of visual working memory arrays has been infrequently examined, so our main goal was to characterize how quickly and to what extent participants can use long-term memory to overcome visual working memory capacity limits when directed to do so intentionally. Indeed, prior work suggests that working memory plays a particularly important role when learning is intentional as opposed to incidental. For example, Unsworth and Engle (2005) found that individual differences in working memory capacity predicted learning in a serial reaction time task in conditions with intentional, but not incidental, learning. To this end, we devised an experimental paradigm to probe the explicit coordination of working and long-term memory. Specifically, we simply instructed participants that the same visual array would repeat for many trials in a row, and that they should use any strategy available to them to improve their performance across repetitions. We predicted that we should initially find that participants' performance is bound by typical working memory capacity limits (i.e., ~3 items), but after many repetitions participants may begin to use long-term memory to augment performance.

Inter-response times as a measure of chunk formation

In addition to examining how accuracy improves with repetitions, we also planned to examine how response latencies may track the formation of new “chunks” in a visuospatial memory task. Here, we used a “whole-report” visual working memory task, where participants are asked to report the color of all memory items. Because participants report multiple items, we may examine not only the number of correctly reported items, but also how quickly participants make individual responses. In particular, we were inspired by prior work measuring inter-response times, defined as the time between pairs of responses as participants make many responses in a row (Anderson & Matessa, 1997; Broadbent, 1975; Browman & O'Connell, 1976; Chase & Ericsson,

¹ Although there is some debate as to how much a redundant verbal code can benefit visual working memory performance, prior work has shown that redundant verbal codes do not greatly boost performance in visual working memory tasks using abstract stimuli. For example, performance has been shown to be equivalent with and without articulatory suppression (Sense et al., 2017). However, emerging work is characterizing how semantically meaningful labels may boost performance by recruiting visual long-term memory (Forsberg et al., 2020; Overkott & Souza, 2021; Souza et al., 2021).

1982; Chase & Simon, 1973; Lovelace & Snodgrass, 1971; Lovelace & Spence, 1972; Murdock & Okada, 1970; Reitman, 1976). During free recall and search through long-term memory, inter-response times tend to increase as the memory set² is exhausted (Bousfield & Sedgewick, 1944; Murdock & Okada, 1970; Rohrer, 1996; Wixted & Rohrer, 1994). In addition to this general slowing over time, clustering of inter-response times can be a useful, quantifiable signature of chunk utilization, whereby “intra-chunk” response times are faster than “inter-chunk” response times (McLean & Gregg, 1967). McLean and Gregg (1967) articulated a framework in which chunks may be formed in three key ways: (1) via *prior knowledge*, (2) via *grouping cues during encoding*, and (3) via *top-down associations* (i.e., new associations formed by attention, rehearsal, or some other process).³ This framework has remained important for later theories of the role of chunking in working memory (e.g., Cowan, 2001).

Inter-response time signatures of chunking have previously been observed when chunks are formed via prior knowledge or during encoding. First, studies examining recall of previously learned sets (e.g., countries of Europe) have found slowing of inter-response times in clusters of three to four items (Broadbent, 1975; Graesser & Mandler, 1978). Second, studies introducing grouping cues during encoding of letter and digit sequences have found a slowing of inter-response times when the recall of a new group begins (Anderson & Matessa, 1997; McLean & Gregg, 1967). Few studies, to date, have looked at inter-response time signatures of chunking when groups are formed via new top-down associations, as we plan to do here (Miller & Unsworth, 2018). However, other putative signatures of chunk utilization have been observed when observers repeatedly recall a word list multiple times in a row (i.e., “multitrial free recall”). Rather than response times, prior work has examined response consistency during multitrial free recall. Namely, when participants encode the same word list multiple times in a row (with the words presented in a randomized order during each list presentation), participants begin to recall the items in a consistent order each time they recall the list (Dunlosky & Salthouse, 1996; Miller & Unsworth, 2018; Sternberg & Tulving, 1977; Tulving, 1962, 1966).

² In long-term memory search studies, the “memory set” refers to the group of items that the participant is trying to recall from memory.

³ Note, here we have reworded McLean and Gregg’s (1967) framework to better reflect today’s vernacular. The original quote (reproduced in Cowan, 2001) from McLean and Gregg (1967) is: “(a) Some stimuli may already form a unit with which *S* is familiar. (b) External punctuation of the stimuli may serve to create groupings of the individual elements. (c) The *S* may monitor his own performance and impose structure by selective attention, rehearsal, or other means.”

In the current study, we were particularly interested in observing how chunks formed via new, top-down associations may benefit performance in a visual memory task. Our whole-report task with explicit repetitions is a novel, visuospatial analogue of the verbal “multitrial free recall” task. The present study will test if classic behavioral signatures of chunk utilization that have been established with verbal memoranda will generalize to visuospatial tasks. Few studies have examined inter-response times in the context of visual memory (with the notable exception of expertise and chess; Chase & Simon, 1973; Reitman, 1976), in part because of the popularity of change detection measures of visual memory that collect only a single response on each trial (e.g., Luck & Vogel, 1997). In the present experiments, we predicted that when participants recruit long-term memory to improve performance beyond typical capacity limits, we would see behavioral signatures of chunking such as “pauses” in the inter-response times and/or increased consistency of response order during recall.

Summary

To preview results, we found a rapid and robust effect of explicit repetitions on performance. Even after only one repetition, participants’ performance exceeded typical capacity limits. By the eighth and final repetition of an array, participants had a modal performance that was perfect (six to eight items correct). In addition, an analysis of inter-response times is consistent with the idea that participants organize their responses by retrieving ‘chunks’ from long-term memory. Together, these findings illustrate how long-term memory may rapidly assist cognition in tasks that overwhelm working memory capacity, and that inter-item response times can be used to track the formation and deployment of chunking strategies in visual memory tasks.

Experiment 1: Six-item arrays

Methods

Participants

Participants were recruited from the University of Chicago and the surrounding community. Participants provided informed consent under procedures approved by the University of Chicago Institutional Review Board. All participants (27 female, 25 male) were 18 years or older ($M = 21.88$ years, $SD = 3.66$ years, range = [18,36]), had self-reported normal color vision and normal or corrected-to-normal visual acuity, and received course credit or cash (\$10/h) for their participation. A total of 52 participants took part in the study. Data from two participants were excluded for failure

to comply with task instructions (i.e., chance-level performance), leaving a total of 50 participants for analysis. The study procedures were not pre-registered, and the sample size was determined by convenience (i.e., data collection up to a conference deadline). With 50 subjects, we would be powered to detect medium within-subjects effects at 90% power (e.g., within-subjects t -test, critical $t(49) = 2.01$, $d_z = .47$; repeated-measures ANOVA with one within-subjects factor (e.g., repetition, eight levels), critical $F(7,343) = 2.04$, Cohen's $f = 0.15$, $\eta^2 = 0.02$) (Faul et al., 2007; Kim, 2016).

Stimuli

Participants were seated in a dimly lit room and viewed a 24-in. BenQ LCD monitor with a 1,920 x 1,080 resolution from a distance of ~67 cm. Stimuli were presented with MATLAB (The MathWorks, Natick, MA, USA) and the Psychophysics toolbox (Brainard, 1997; Kleiner et al., 2007). A fixed set of nine highly discriminable colors were used for the colored square stimuli in all three memory tasks (red: 255 0 0; green: 0 255 0; blue: 0 0 255; yellow: 255 255 0; magenta: 255 0 255; cyan: 0 255 255; white: 255 255 255; black: 1 1 1; orange: 255 128 0), and colors were always chosen without replacement for each memory array. Throughout each task, a black fixation dot was drawn at the center of the screen (radius = 6 pixels, 0.14°) and stimuli were presented on a medium-gray background (RGB = 85 85 85).

Discrete whole-report task with repetitions A total of 30 unique arrays were generated by picking six semi-random locations and assigning a unique color (drawn from the set of nine possible colors) to each location. The locations were semi-random in that they were chosen with some constraints, such that items were separated by a minimum distance of 36 pixels ($\sim 0.9^\circ$ of visual angle) and were split evenly across the left and right hemifields. Each colored square had a diameter of 72 pixels ($\sim 1.7^\circ$) and the possible locations were in a portion of the screen centered on fixation and subtending 1,066 x 600 pixels (7.1° above/below fixation and 12.6° left/right of fixation).

Surprise long-term memory recognition task For the long-term memory recognition task, we showed participants a total of 60 arrays (30 old, 30 new). The old arrays were identical to those used in the whole-report task. The new arrays were randomly generated with the same size, color, and location constraints as in the working memory task.

Color change detection task On each trial, a new array containing four, six, or eight colored squares was drawn. The stimuli were the same size and drawn in the same nine colors

as the whole-report task, and the same location constraints were used.

Procedures

Participants completed a discrete whole-report task with array repetitions, a surprise long-term memory recognition task for the arrays presented in the whole-report task, and a color change detection task. These three tasks were presented in a fixed order for all participants.

Repeated-arrays working memory task Participants completed a variant of a discrete whole-report working memory task (Adam et al., 2015; Huang, 2010) in which arrays repeated eight trials in a row. On each trial, participants saw a briefly presented array of six colored squares (150 ms). After a short delay (1 s), participants reported the colors of the squares. A 3 x 3 grid of possible color choices appeared at each location, and participants selected the color that belonged at each response grid location. Participants were required to make a response to all six squares before they could advance to the next trial. After the last response was made, the next trial began after an inter-trial interval of 1 s. Critically, the same array was repeated eight times in a row. On the first trial of a set of repetitions, a new configuration of square colors and locations was randomly chosen. This array was then used for the next seven working memory trials in a row (i.e., trials 1–8 were array #1, trials 9–16 were array #2, etc). Participants were given explicit instructions that each unique array would be repeated for eight trials in a row, and that they should try to improve their performance across the eight repetitions. Participants completed a total of 240 working memory trials (eight trials each of 30 unique arrays).

Old-new recognition task After completing the repeated arrays working memory task, participants completed an old-new recognition task for the 30 arrays that were used in the repeated arrays working memory task. Participants were not informed beforehand that they would be tested on their long-term memory of the arrays in the previous task. On each trial, participants viewed an array of colored squares. On half of the trials, the participants were shown an old array (an array that was previously seen in the discrete whole-report task). On the other half of the trials, the participants were shown a new, randomly generated array with the same stimulus constraints (i.e., six colored squares drawn at new random locations). Participants reported via keypress if they thought the array was “old” (“Z” key) or “new” (“P” key) and they reported their confidence about the decision on a 5-point scale (using the number keys 1–5 on the keyboard). All responses were unsped.

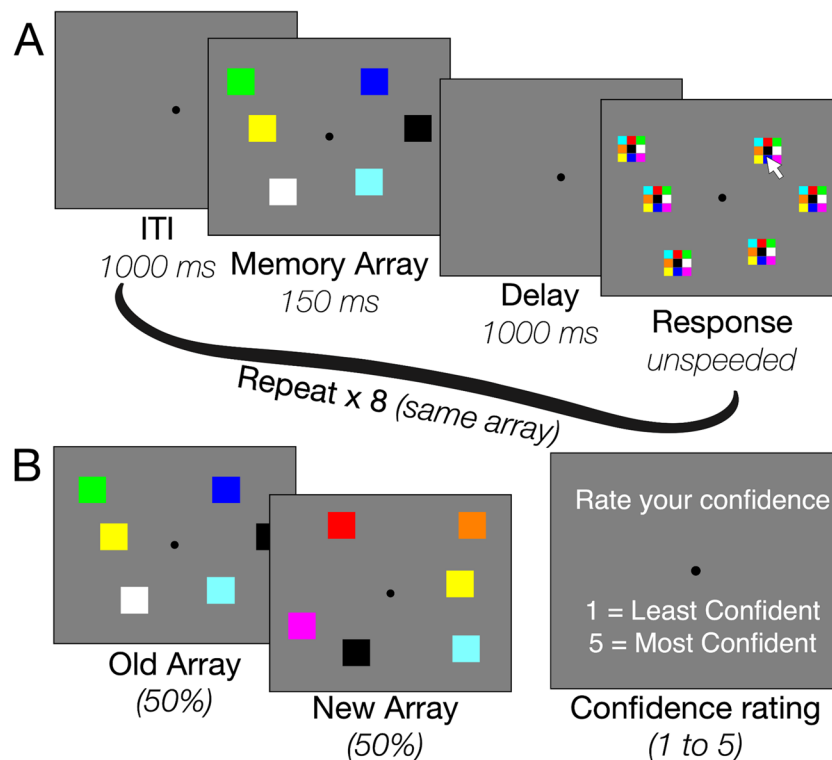


Fig. 1 (A) Repeated arrays working memory task. On each trial, participants remembered an array of colored squares across a blank delay. At test, participants used the mouse to report which color was presented at each of the locations. For example, if the upper right square was blue, the participant clicked the blue portion of the mask. Each memory array was repeated for eight trials in a row. Partici-

pants were instructed to try to improve their performance across the eight repetitions. (B) Old-new recognition task. After completing the repeated arrays working memory task, participants were given an old-new recognition task (50% old arrays from the previous task, 50% new arrays). Participants reported whether they thought the array was old or new, as well as their confidence in their decision (5-point scale)

Color change detection task To assess baseline working memory performance with an independent task, we used a standard color change detection task (Luck & Vogel, 1997). On each trial, participants saw a briefly presented array of four, six, or eight colored squares (150 ms), and remembered the colors of the squares across a blank delay (1 s). At test, a memory probe was shown at one of the squares' locations. On half of the trials ("same" trials), the probe was the same color as the remembered item at the same location. On the other half of trials ("different" trials), the probe was a different color. Participants responded via keypress whether they thought the probe square was the same ("Z" key) or different ("/" key) from the remembered color of the square presented at the probe's location. Participants completed a total of 180 trials of the color change detection task (60 trials per set size).

Analysis

Analyses were performed using MATLAB 2018a (The MathWorks) and Python 3.9.7 (conda 4.12.0). Data from

the raw .mat files were processed in MATLAB and converted into aggregate .csv files for the main analyses in Python. Key open source packages for Python analyses include Jupyter (Kluyver et al., 2016), pandas (McKinney, 2010), seaborn (Waskom, 2021), pingouin (Vallat, 2018), and pymer4 (Jolly, 2018) Fig. 1.

Results

Performance rapidly increased across array repetitions

To characterize how performance changed as a function of repetition, we first analyzed mean performance (Fig. 2A). In the whole-report task, performance is quantified as the number of locations for which participants correctly recalled the item's color, and this value ranges from 0 to 6 on each trial. The first time the participants saw an array (Repetition 1), mean performance was in line with typical estimates of working memory capacity ($M = 2.79$ items correct, $SD = 0.45$). Mean performance

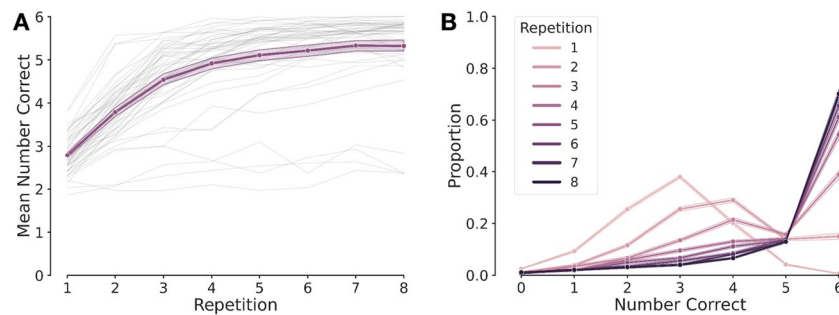


Fig. 2 Working memory performance as a function of array repetition in Experiment 1. **(A)** Mean number correct as a function of repetition. The number of correctly recalled items increased dramatically from the first repetition to the later repetitions. The purple line indicates the mean performance increase; thin gray lines depict individual subjects; shaded error bars represent 68% confidence intervals (approximately equivalent to standard error). **(B)** Distribution of performance

significantly increased across repetitions, $F(7,343) = 330.6$, $p < 1 \times 10^{-45}$, $\eta_p^2 = 0.871$.⁴ By the final repetition, participants' performance was near ceiling and had nearly doubled from the first repetition ($M = 5.32$, $SD = 0.91$). To quantify the rate of performance improvement on average, we calculated difference scores for adjacent repetitions (e.g., Repetition 2–1, Repetition 3–2, etc.). On average, participants' performance improved by 0.36 items per repetition ($SD = .11$), with faster learning across the first four repetitions ($M = .71$, $SD = .24$) compared to the last four repetitions ($M = .10$, $SD = .09$), $t(49) = 16.4$, $p < 1 \times 10^{-20}$. In Experiment 1, the ceiling was six items correct. As such, the slowing of learning at later repetitions may have been driven by participants hitting the performance ceiling for the task.

In addition to looking at mean performance for each repetition, we also looked at the distribution of performance outcomes (Fig. 2B). One notable aspect of the performance distributions is the increase in the number of trials where participants correctly recalled six out of six items. In a typical whole-report working memory task, participants rarely get six items correct, and these rare "perfect" trials can be explained by guessing inflation (i.e., participants never really store six items, but they sometimes get lucky and get six correct by chance because they are required to make a response to every item; see Adam et al., 2015). The first time participants saw an array (Repetition 1), we found a similar pattern of performance. There was a strong modal tendency toward getting three items correct, and participants very rarely got all six items correct ($M = 0.6\%$, $SD = 1.87\%$). As early as the second encounter with the

outcomes as a function of repetition. On the first repetition, participants' performance resembled typical working memory performance (mode = three items correct; few or no responses with six items correct). By the third repetition, the modal tendency was six out of six correct, far exceeding typical working memory capacity limits. Shaded error bars indicate 68% confidence intervals

array (Repetition 2), the number of perfect trials increased 25-fold (from 0.6% to 15%). By the final encounter with the array (Repetition 8), the modal tendency was six out of six correct ($M = 65.5\%$, $SD = 24.6\%$).

Inter-response times and chunk utilization

Prior work has hypothesized that inter-response times can be used as a signature of retrieving a new chunk from long-term memory (e.g., Broadbent, 1975). Inter-response times are calculated as the time in between successive responses, and a long pause may indicate that a participant is engaging in planning for the next series of responses and/or retrieving information from long-term memory. The inter-response times are shown in Fig. 3A. The response time was longest the first time participants saw an array (Repetition 1), and the successive responses became quicker. Starting on the second repetition of the array (Repetition 2), we observed a marked slowing at the fourth response. A two-way repeated-measures ANOVA with within-subjects factors Response Number and Repetition confirmed that there was a significant interaction of Response Number and Repetition on inter-response times, $F(35,1715) = 6.34$, $p < .001$, $\eta_p^2 = .11$. To better understand the effect of repetition on inter-response times, we conducted separate one-way repeated-measures ANOVAs with factor Repetition separately for each response. To visualize the meaning of these tests, the data for each response are replotted in separate subplots in Fig. 3B.

We found significant effects of repetition on inter-response times for the first response, $F(7,343) = 8.81$, $p < .001$, $\eta_p^2 = .15$, for the fourth response, $F(7,343) = 5.27$, $p < .001$, $\eta_p^2 = 0.10$, for the fifth response $F(7,343) = 8.12$, $p < .001$, $\eta_p^2 = 0.14$, and for the sixth response,

⁴ Greenhouse-Geisser corrected p-values are reported when the assumption of sphericity is violated.

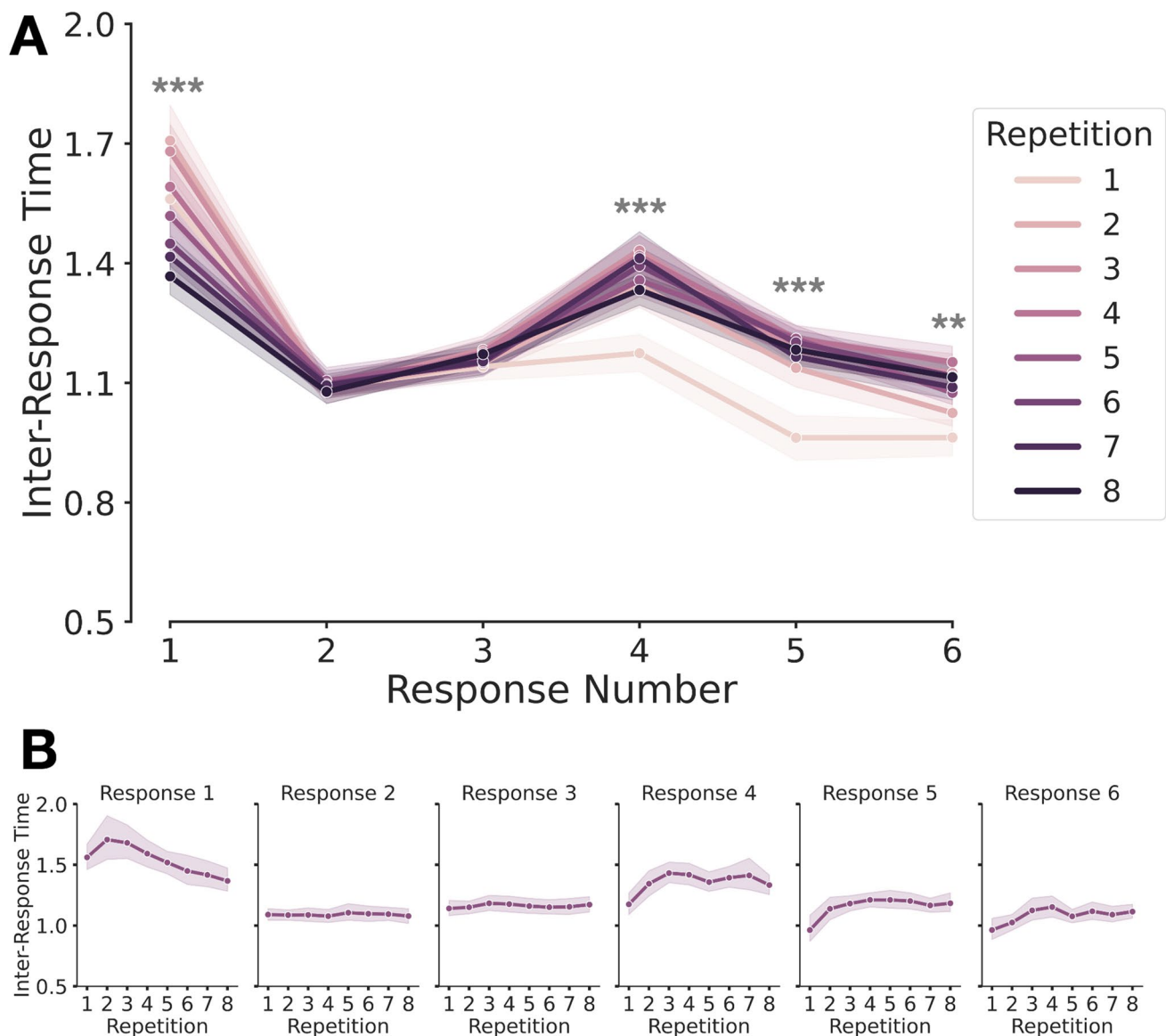


Fig. 3 Inter-response times as a function of response number and repetition number in Experiment 1. **(A)** Inter-response times plotted with response number on the X-axis and separate lines for repetition. Asterisks indicate that there was a significant effect of repetition on inter-response times for a one-way ANOVA for that response number

(uncorrected; ** $p < .01$, *** $p < .001$). **(B)** Inter-response times re-plotted with repetition on the X-axis and separate subplots for each response. Error bars represent 68% confidence intervals (approximately equivalent to 1 standard error of the mean)

$F(7,343) = 4.94$, $p < .001$, $\eta_p^2 = 0.09$. In contrast, there was no effect of repetition on inter-response times for Response 2 ($p = .92$) and Response 3 ($p = .74$). This pattern of response times would be consistent with a chunking strategy where participants formed an initial chunk of three items on Repetition 1 that they used throughout the subsequent repetitions. However, starting at Repetition 2, participants appear to have become more efficient at using their already formed chunk (faster response times for response 1 with repetition) and devote extra time during the fourth response to form and recall a second chunk

of three items (slower response times for response 4 with repetition).

In addition to response times, we also examined whether participants recalled items in a consistent order by computing transition probabilities between all pairs of items. A transition probability of 100% would indicate that participants reported a pair of items in the same order for all eight repetitions of the array. We found that participants reported items in an order that is more consistent than would be expected by chance ($p < .001$), with the highest two transition probabilities exceeding 90% (see Online Supplemental Material

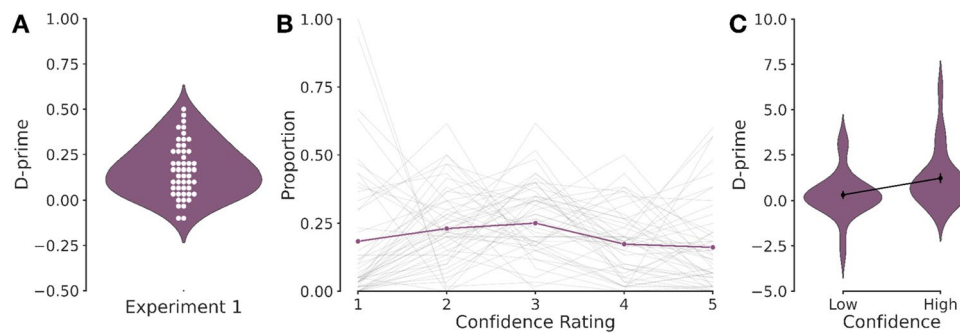


Fig. 4 Performance in the surprise recognition task. **(A)** Overall d' : Participants were significantly above chance ($d' > 0$) at distinguishing old arrays from new, randomly generated arrays. Purple shaded outline depicts the distribution of the data. White dots show individual subjects' scores. **(B)** Histogram of confidence scores (1 = lowest confidence, 5 = highest confidence). The purple line represents the mean proportion that a confidence rating was used, error bars indicate 68%

(OSM) Analysis S2; Fig. S1A). The empirical pattern that we observed is consistent with an account in which participants formed links between the first three items starting on Repetition 1 (i.e., two transition probabilities: Item 1- \rightarrow 2 and Item 2- \rightarrow 3), and then developed a consistent response order for the remaining items during later repetitions. Furthermore, participants' response order was more consistent for the later repetitions of the array (Repetitions 5–8) than for the early repetitions of the array (Repetitions 1–4; Fig. S1B (OSM)). This is consistent with the notion that participants first successfully remembered a few items, and then added in more items as the array was repeated (see also Fig. S2A (OSM)).

Successful recognition of repeated arrays in an old-new recognition task

We hypothesized that participants were able to exceed typical working memory capacity limits by rapidly recruiting long-term memory. Given this hypothesis, we next tested whether participants could reliably distinguish learned arrays from novel arrays in the old-new recognition task. We quantified long-term memory performance as d' (Fig. 4A), calculated as the normalized difference between hit rate and false alarm rate ($d' = z(H) - z(FA)$ where $z()$ is the z -transform) (e.g., Banks, 1970). Overall recognition performance was $d' = 0.45$ ($SD = 0.40$), and this was significantly above chance $t(49) = 7.94$, $p < .001$ (one-tailed t -test⁵). For correlations between individual differences in

confidence intervals (approximately equivalent to standard error). Gray lines depict individual subject histograms. **(C)** Performance split by low confidence (< 3) versus high confidence (> 3) for the $n = 41$ subjects with sufficient trial counts in both confidence bins. Purple shaded outlines depict the distribution of the data. Black line and error bars represent the mean scores and 68% confidence intervals

change detection performance, learning rate in the whole-report task, and recognition memory performance, see Analysis S1 (OSM).

We next examined whether recognition memory performance varied as a function of confidence. Figure 4B shows the distribution of confidence ratings. Overall, the distribution of confidence scores was fairly even; a one-way repeated-measures ANOVA revealed no difference in the frequency with which participants used each confidence level ($p = .12$). To examine recognition memory performance as a function of confidence, we divided trials into “low-confidence” (< 3) and “high-confidence” (> 3) bins. Given the total number of trials available for analysis (60), not all participants had sufficient numbers of trials to determine d' for both the low- and high- confidence bins (i.e., 0 hit or false alarm trials in a given confidence bin). After excluding subjects with insufficient data, there were 41 subjects available for a within-subjects analysis. A paired t -test revealed a significant effect of confidence on memory performance, $t(40) = 2.72$, $p = .009$, where memory performance was significantly better for high-confidence trials ($M = 1.23$, $SD = 1.56$) compared to low-confidence trials ($M = 0.32$, $SD = 1.24$)⁶ (Fig. 4C). Memory performance was significantly above chance for high-confidence trials ($p < .001$) but not for low-confidence trials ($p = .06$).

⁵ Note, a one-tailed t -test was chosen for this comparison, because we would not expect memory performance to be meaningfully below chance. All statistical tests are two-tailed unless otherwise noted.

⁶ Note, an independent t -test including all conditions with enough trials in at least one of the two confidence bins yielded similar results ($n = 43$ high confidence; $n = 48$ low confidence), $t(89) = 3.15$, $p = .002$.

Experiment 2: Eight-item arrays

To replicate our results, we performed a second experiment that closely parallels Experiment 1. The only key change that we made was to increase the set size to eight items instead of six items for each array. By raising the set size, we increased the potential performance ceiling even further beyond typical capacity limits of three to four items.

Methods

Participants

An additional 60 participants were recruited from the University of Chicago and the surrounding community (32 = female, 28 = male; mean age = 21.29 years, $SD = 3.20$, range = [18,35]). A total of three participants were used in pilot sessions to test the task code and the length of the session (e.g., different numbers of unique arrays and repetitions); five participants were excluded because of incomplete datasets (two computer crashes leading to lost data; three participants did not complete all three of the memory tasks), leaving a total of 52 participants for analysis. This study was designed as a close replication of Experiment 1, but the study procedures were not formally pre-registered; the sample size was chosen to approximately match Experiment 1.

Stimuli and procedures

The stimuli and procedures in Experiment 2 were nearly identical to those in Experiment 1 with the following changes. First, in the repeated working memory task and old-new recognition task, each memory array contained eight squares (rather than six squares) and participants saw a total of 27 unique arrays in the repeated working memory task (rather than 30). Second, participants were given a short survey at the end of the task in which they answered a free-response question about any strategies used and made numerical ratings of the number of items they thought they got correct on average and their overall feelings of effort, boredom, drowsiness, enjoyment, frustration, motivation, challenge, and distraction during the experiment.

Results

Replication of key results: Rapid learning and above-chance recognition

We replicated the main findings that working memory performance rapidly improved across repetitions and that participants could reliably distinguish learned arrays from novel arrays in a later test. Participants' overall performance

rapidly increased across the eight repetitions, $F(7,357) = 213.8$, $p < .001$, $\eta_p^2 = .81$ (Fig. 5A). After the first encounter with a new array (Repetition 1), performance was in line with typical estimates of working memory capacity ($M = 2.77$ items, $SD = 0.51$). By the last repetition (Repetition 8), overall performance had more than doubled ($M = 6.03$ items, $SD = 1.54$). Figure 5B shows a histogram of trial outcomes (i.e., the proportion of trials where participants got 0 through eight items correct). We again found that performance resembled typical results from whole-report tasks on the first repetition, with a modal tendency of two to three items and few or no trials with perfect performance (mean proportion of trials with eight out of eight correct on Repetition 1 = 0.000%, $SD = 0.000$). By the final repetition, the modal tendency was eight out of eight correct ($M = 40.6\%$ of trials, $SD = 26.7\%$). We again calculated difference scores for adjacent repetitions (e.g., Repetition 2–1, Repetition 3–2, etc.). On average, participants' performance improved by 0.47 items per repetition ($SD = .19$), with faster learning across the first four repetitions ($M = .73$, $SD = .39$) compared to the last four repetitions ($M = .26$, $SD = .11$), $t(51) = 8.76$, $p < 1 \times 10^{-11}$.

In the long-term memory recognition task, participants were above chance at distinguishing old arrays from new arrays (Fig. 5C), $d' = 0.38$ ($SD = 0.41$), $t(51) = 6.65$, $p < .001$ (one-tailed t-test). Unlike Experiment 1, participants in Experiment 2 used the confidence scores at unequal rates, $F(4,204) = 20.9$, $p < .001$, $\eta_p^2 = .29$. After excluding participants with insufficient data to quantify d' for high- (> 3) and low- (< 3) confidence trials, we still had 41 participants remaining for the analysis of d' as a function of confidence. We again found that d' was better for high-confidence trials ($d' = 0.96$, $SD = 2.09$) compared to low-confidence trials ($d' = 0.09$, $SD = 1.58$), $t(40) = 2.13$, $p = .039$ ⁷. Overall, memory performance was significantly above chance for high-confidence ($p = .003$) but not for low-confidence ($p = .36$) trials.

Flexible chunking strategies for different set sizes

We again found signatures of chunking strategies that changed as a function of repetition when analyzing the inter-response times. Here, however, we found that participants grouped their responses into sets of two rather than into sets of 3 (Fig. 6A). As in Experiment 1, we ran an initial two-way repeated-measures ANOVA with within-subjects factors Response Number and Repetition confirmed that there was a significant interaction of

⁷ Likewise, we again found a consistent result when we included all possible conditions with an independent t-test ($n = 45$ high confidence; $n = 48$ low confidence), $t(91) = 2.46$, $p = .016$.

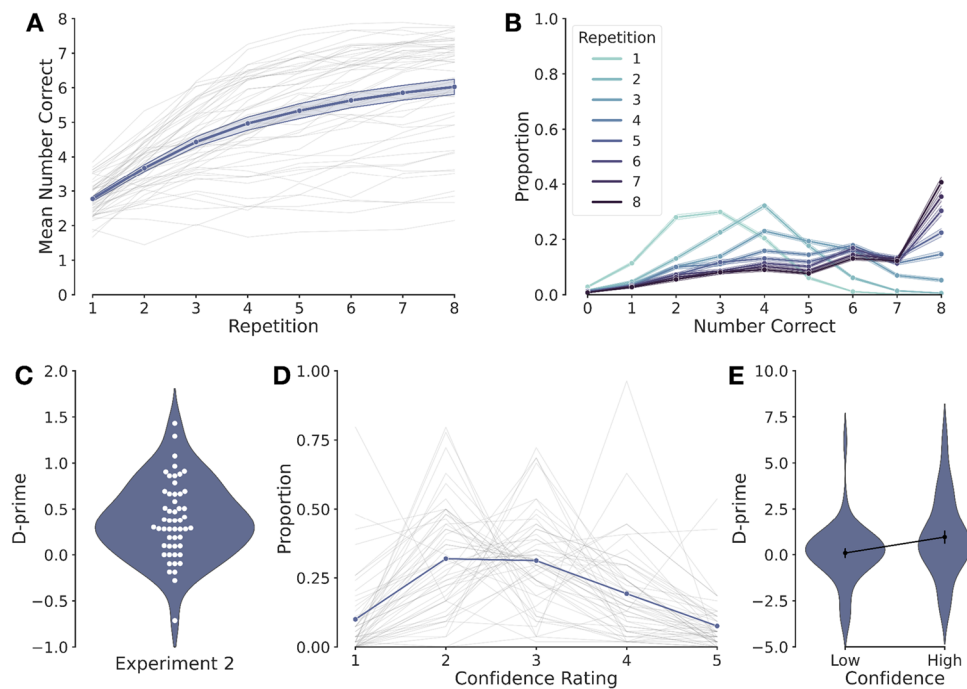


Fig. 5 Working and long-term memory performance in Experiment 2. **(A)** Improvement in mean performance across repetitions. The blue line represents average performance, with shaded error bars indicating 68% confidence intervals (approximately equivalent to standard error). Thin gray lines represent individual participants. **(B)** Histogram of trial outcomes, from zero to eight items correctly recalled. **(C)** Overall recognition memory performance (d'). The blue violin shows the shape of the distribution; the white dots show individual participant values. **(D)** Distribution of confidence ratings from 1 (lowest confidence) to 5 (highest confidence). The blue line represents the average distribution; the thin gray lines show individual

participants' data. **(E)** Recognition memory performance split by high (> 3) and low (< 3) confidence. The blue violins show the distribution of the data; the black line with error bars show the mean d' values with 68% confidence intervals (We used version 0.11.2 of the plotting package seaborn, which only allows one error bar type (confidence interval). We wanted to plot approximately 1 standard error of the mean (SEM), as this is the error bar type we typically use in other published work. The 95% confidence intervals are approximately equal to 1.96 SEMs, and 68% confidence intervals are approximately equal to 1 SEM. Note, the error bars are for visualization purposes only, and do not directly impact the interpretation of results)

Response Number and Repetition on inter-response times, $F(49,2499) = 8.14$, $p < .001$, $\eta_p^2 = .14$. To understand this interaction, we ran follow-up one-way ANOVAs with factor Repetition for each response, and we replotted the data in Fig. 6B. There was a modest effect of repetition on inter-response times for responses 1, 3, and 4 ($p < .05$, $\eta_p^2 = .06$), and no effect of repetition at response 2. There was a robust effect of repetition on inter-responses in the last four repetitions ($p < .001$), which was particularly pronounced for responses 5 through 7 ($\eta_p^2 = .22 - .34$). In sum, it seems that participants flexibly adapted their chunking strategy to use groups of two rather than three. Starting at repetition 2, we observed a marked slowing of response times for the fifth response, suggesting that participants began forming a third chunk of two items after just one encounter with the array. In contrast, we did not see significant slowing of the seventh response until repetition 3, suggesting that participants tended to attempt one new chunk formation with each repetition. We repeated the supplemental transition probability analysis, and found

results consistent with Experiment 1 (Figs. S2B and S3 (OSM)).

Qualitative survey

In a post-experiment questionnaire, we asked participants to describe the strategy that they used to complete the task, whether their strategy changed across repetitions, metacognitive estimates of the number of items they got correct on the first and last repetition, as well as numerical ratings of effort, boredom, drowsiness, enjoyment, frustration, motivation, challenge, and distraction. Note, the first three participants that were run did not complete the survey, leaving $N = 49$ participants for the survey analyses.

First, we provide an overview of participants' open-ended responses about strategy use. A total of three raters (two authors) independently rated each survey response for the presence or absence of seven strategy categories. Raters were allowed to endorse more than one category (i.e., for all seven strategies, the rater decided if it was

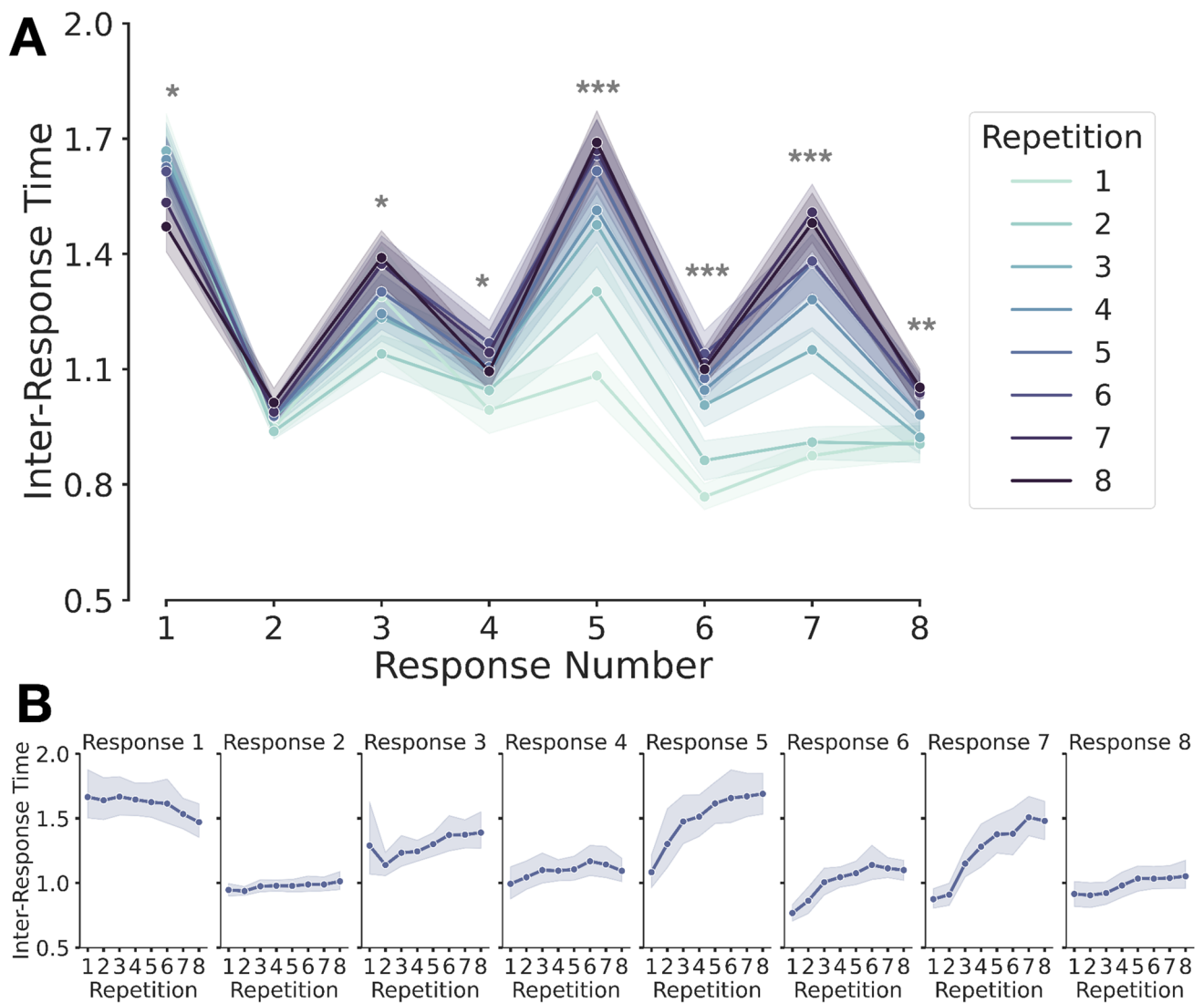


Fig. 6 Inter-response times as a function of response number and repetition number in Experiment 2. **(A)** Inter-response times plotted with response number on the X-axis and separate lines for repetition. Asterisks indicate that there was a significant effect of repetition on inter-response times for a one-way ANOVA for that response num-

ber (* $p < .05$, ** $p < .01$, *** $p < .001$). **(B)** Inter-response times replotted with repetition on the X-axis and separate sub-plots for each response. Error bars represent 68% confidence intervals (approximately equivalent to 1 standard error of the mean)

present or absent). The strategy categories included (1) spatial grouping (e.g., “I would focus on remembering two squares at a time”), (2) overt or covert verbal rehearsal (e.g., “I would say the colors out loud to try and memorize the colors”), (3) forming spatial paths (e.g., “I sought to remember squares from left to right then in a clockwise directed oval) (4) forming verbal paths (e.g., “I also memorized the word sequence, not really the color sequence”), (5) salience (e.g., “...first two squares which caught my attention first”), (6) semantic (e.g., “I paired colors that were associated with each other. For example, red + white + blue = american flag, orange + black = Halloween”) and

(7) random (e.g., “Majority of the time I just randomly picked squares to remember and it did not work”).

Inter-rater agreement was overall good, with mean agreement of 88.4% across all strategies (minimum: 78.23% for spatial grouping; maximum: 98.64% for salience). To quantify how common each strategy was, we coded each strategy as present for an individual when at least two out of three reviewers agreed. This revealed that the most commonly used strategies were visual grouping ($N = 34$) and verbal rehearsal ($N = 18$) followed by: forming a spatial path ($N = 11$), forming verbal paths ($N = 5$), salience ($N = 4$), semantic ($N = 3$) and random ($N = 1$). Many participants’

responses consisted of a mixture of two or more of these strategies. For example, some participants reported using a mixture of visual grouping and overt or covert verbal rehearsal ($N = 10$, e.g., “I split the arrays into pairs of squares; I also repeated the first four squares’ colors in my head verbally many times”). Task performance is plotted separately as a function of strategy in Fig. S4 (OSM).

Approximately equal numbers of participants said their strategy changed ($N = 25$) versus did not change ($N = 24$) from the first to the last repetition. Note, we had intended for this question to reflect changes in strategy across the eight repetitions for each individual array (i.e., whether participants used a different strategy the first time versus the last time they saw a particular array). However, almost all the participants who answered “yes” seem to have interpreted this question as asking whether they changed their strategy from early in the session to later in the session. For example, some participants reported global changes to their strategy over the session, whereby they initially did not have a good strategy for performing the task: “developed strategy more as I went along”; “I tried visualizing the entire screen in the beginning but it was too hard to take in all the squares at one time”; and “strategy varied with motivation.” Other participants reported some fine-grained change to their strategy, or a shift from a more visual to a more verbal strategy as the task progressed: “towards halfway point I changed light blue to teal in my mantra to avoid confusing it with blue square”; “I started to use words instead of trying to memorize colors directly.”

As a group, participants’ metacognitive estimates of the number of items they correctly reported were well-calibrated. Participants estimated that they stored on average 2.31 items ($SD = .89$) on the first repetition and 6.18 items ($SD = 1.76$) on the last repetition. In comparison, the ground truth numbers were 2.79 on the first repetition and 6.10 on the last repetition. Note, however, this group-level similarity in estimated and actual performance does not guarantee metacognitive accuracy at the level of individuals (e.g., it may be that no participants were accurate, and those that over- and under-estimated performance canceled each other out). To test the specificity of these estimates for each individual, we also computed correlations and a difference metric. For the correlation, we correlated each participant’s mean self-estimate with their mean performance. For Repetition 1, participants’ average performance correlated with their self-estimate, $r = .55$, $p < 1 \times 10^{-4}$,

slope = .94, intercept = $-.32$. For Repetition 8, participants’ average performance also correlated with their self-estimate, $r = .76$, $p < 1 \times 10^{-9}$, slope = .89, intercept = 0.79.⁸ For the difference metric, we calculated the absolute value of the difference between each participant’s performance and their self-estimate. For Repetition 1, the mean absolute difference between actual and self-estimated performance was 0.69 items ($SD = .56$). For Repetition 8, the mean absolute difference between actual and self-estimated performance was 0.82 items ($SD = .81$).

Finally, we quantified the subjective ratings obtained from participants about their state of mind during the task. With 1 being “minimum level of X” and 5 being “maximum level of X”, participants’ reported having a level of effort of 3.94 ($SD = 0.75$), a level of boredom of 2.92 ($SD = 0.93$), a level of drowsiness of 2.63 ($SD = 1.24$), a level of enjoyment of 2.49 ($SD = 0.87$), and a level of frustration of 2.31 ($SD = 1.23$). When asked how motivated they felt to do their best (1 = not at all motivated, 5 = extremely motivated), participants reported a rating of 3.63 ($SD = 0.76$)⁹. When asked how challenging they found the experiment (1 = not at all challenging, 5 = extremely challenging), participants reported a rating of 3.63 ($SD = 1.05$). Finally, when asked how distracted participants felt by thoughts about their own life while doing the experiment (1 = almost never, 5 = extremely frequently), participants reported a rating of 2.63 ($SD = 1.11$).

Discussion

To accomplish our goals, we frequently need to use working and long-term memory in tandem. However, in scientific studies of memory, we typically try to study memory systems in isolation. In two behavioral experiments, we used explicit repetitions of memory arrays to study the interaction of working and long-term memory. Like many studies of working memory, here we started with stimuli that were abstract and devoid of any pre-existing long-term memory associations (i.e., unique, arbitrary pairings of colors and locations). To allow for the recruitment of long-term memory, we explicitly repeated each array eight times in a row. Using this method, we were able to watch the interaction of visual working- and long-term memory unfold over time. We found rapid improvement of working memory performance to levels far beyond typical visual working memory capacity limits. On average, we found that working memory performance increased at a rate of around 0.4 items per repetition. However, this improvement in working memory

⁸ Here, we assessed participants’ self-estimates of performance after they had already completed the task (post-diction). Prior work on memory self-efficacy has shown that self-estimates made before ever experiencing a task (predictions) would be less strongly correlated with performance compared to post-dictions (Beaudoin & Desrichard, 2011; Hertzog et al., 1990, 1994; West et al., 1996).

⁹ One additional participant was missing a response for the motivation item ($N = 48$ for this item).

performance over time was non-linear, with participants showing a rapid increase in performance over the first few repetitions (~0.72 items per repetition for Repetitions 1–4) followed by a much slower increase (~0.18 items per repetition for Repetitions 5–8). After only a few repetitions, modal performance was perfect for six and eight item arrays that are typically far beyond working memory capacity (Adam et al., 2015). The rapid recruitment of long-term memory observed here is consistent with prior EEG results demonstrating that participants can flexibly hand off a visual search template from working to long-term memory, and that they can flexibly recruit one or both memory systems depending on task demands (Carlisle et al., 2011; Reinhart & Woodman, 2014).

Maximum achieved performance across the two experiments

In both experiments, we found that participants approximately doubled their initial performance as the array repeated, from 2.8 items on the first repetition to 5.3+ items on the eighth repetition (Exp. 1 = 5.32, Exp. 2 = 6.03). Notably, however, the ceiling was also different in the two experiments, with a maximum number of six correct possible in Experiment 1 and a maximum number of eight correct possible in Experiment 2. In Experiment 1, we speculated that we may not have been able to observe further improvement to performance because of a ceiling effect. This potential ceiling effect may have artificially slowed the observed learning as participants approached the ceiling. However, when we raised the set size to eight items in Experiment 2, group-averaged learning rates were not similarly constrained by the ceiling, since average maximum performance was still around six items. Given this finding, the maximum performance that we observed in Experiment 1 may not have been determined entirely by the ceiling, but may instead reflect the rate of learning that is possible across eight repetitions when using highly similar visuospatial arrays as memoranda.

Although some participants achieved ceiling performance in Experiment 2, there was a wide spread of individual differences such that many participants failed to reach the ceiling. In particular, one factor that may have slowed learning in both experiments was the build-up of interference. Specifically, given a limited set of only nine possible colors, there would have been a high degree of perceptual similarity

for the ~30 learned arrays. For example, in Experiment 2 (27 unique arrays), eight out of nine possible colors were used for every array and only the specific color-location pairings distinguished the arrays from each other. Indeed, although participants performed above chance on the old-new recognition task in both experiments, recognition memory performance was fairly low (Exp 1 $d' = .45$, Exp 2 $d' = .38$). This recognition memory performance is lower than has been observed when participants learn a single critical repeated array (Musfeld et al., 2023a, 2023b; Souza & Oberauer, 2022). For example, Musfeld et al. (2023b) found that recognition memory performance for the single learned array was very high ($d' \approx 1.2$ for the “recall one” learning condition [Rec(1)-Rec(1)], $d' \approx 2.4$ for the “recall all” learning condition [Rec(6)-Rec(6)]).¹⁰ Thus, we think that the high degree of overlap between the ~30 learned arrays in our study contributed to the fairly low recognition memory performance we observed. Future work manipulating the degree of similarity between learned arrays (e.g., by adding context) will be useful for characterizing how the buildup of interference manifests during the interaction of working- and long-term memory.

Inter-response times show clustering consistent with chunk formation

Finally, we found support that inter-item response times track the formation of new chunks in visuospatial memory tasks. The inter-response time data support an account whereby participants initially encode a single chunk of two to three items, and then encode new chunks with subsequent repetitions. In Experiment 1, participants initially showed a notable slowing only on the first response; starting on the second repetition, we observed a second slowing at the fourth response, consistent with the creation and retrieval of a second chunk of three items. In Experiment 2, participants instead grouped pairs of items, but they again showed a formation of additional chunks only for later repetitions. Note, because the stimuli and the responses were at the same spatial locations on each repetition, it is also possible that rote learning for the motor sequence performed while making responses contributed to our learning effects (Carlson et al., 1993). Future work is needed to quantify the relative contribution of “real” long-term memory chunks from motor-planning chunks in inter-response time data like ours.

Limitations and future directions

Future work is needed to understand the specific role of explicit repetitions from other aspects of the task that may aid the recruitment of long-term memory. For example, here we chose to use a whole-report task, which requires participants to make a response for every item in the array.

¹⁰ Note, d' values are approximate. Values were extracted from Fig. S1 of the Supplementary Materials of Musfeld et al. (2023b) using WebPlotDigitizer (Rohatgi, 2022). Specifically, the authors reported the probability that the participant responded “old” for the critical Hebb array (hits) versus for a randomly generated “new” array (false alarms). We calculated d' from these two values to make them comparable to how we quantified long-term memory performance.

Whole-report tasks provide a lot of retrieval practice, and they may also encourage participants to effortfully encode items because they expect a difficult test. Prior work has found that these particular task factors are key for observing significant learning across incidental repetitions of visual arrays (Musfeld et al., 2023a, 2023b; Souza & Oberauer, 2022). With both explicit repetitions and a whole-report task, we found rapid improvement in performance. Interestingly, this improvement mirrors analyses of individuals' learning curves conducted in work by Musfeld and colleagues (2023a) using incidental repetitions of a single critical repeated array (the "Hebb array") and a whole-report task. When analyzed at a group level, incidental learning of the Hebb array appeared to be gradual. However, when analyzed at the individual level, learning appeared to follow a two-step process, with a period of no learning followed by a period of rapid improvement. Critically, the onset of the rapid learning period was related to participants explicitly recognizing that the critical Hebb array was being repeated. In sum, prior results nicely mirror our findings about the ability of explicit repetitions to shape learning (Huang & Awh, 2018; Musfeld et al., 2023a, 2023b; Ngiam, Brissenden, et al., 2019a, 2019b). Future work will be useful to directly compare (1) the onset of learning in an incidental versus explicitly instructed learning context, and (2) making one response versus multiple responses during incidental and explicit learning (Heinen et al., 2022; Musfeld et al., 2023b).

A number of factors could lead to the performance improvement that we observed, including encoding time, elaborative encoding, and retrieval practice. By the end of the eight repetitions, participants had viewed the stimuli for a total of 1,200 ms (150 ms x 8), and they encoded, remembered, and retrieved the items eight different times. Based on prior work, we think that the amount of encoding time, alone, is unlikely to explain the improvements that we observed. For example, in prior behavioral and EEG work using visuospatial working memory tasks, performance was no better when participants were given 200 versus 2,000 ms to encode a visuospatial array (Brady et al., 2016; Tsubomi et al., 2013)¹¹. In contrast, there is ample evidence that retrieval practice and testing robustly improve performance in typical free recall tasks (e.g., Rowland, 2014). Based on a qualitative examination of the survey that we administered, we found that while participants reported using many different strategies, the majority of participants used a spatial grouping strategy (remembering pairs of items), a verbal

rehearsal strategy (overtly or covertly repeating color names) or a combination of these two strategies. However, few participants reported using a strategy that relies on elaborative encoding of the semantic associations of the colored squares (e.g., flag colors). Taken together, we would speculate that retrieval practice and chunking, as opposed to encoding time and semantic associations, most strongly contributed to the performance improvements in our data. In future work, it would be useful to directly manipulate the strategy assigned to participants, to quantify how semantic associations may further boost the rate of learning when participants are instructed to use elaborative versus rote rehearsal strategies (e.g., Craik & Lockhart, 1972).

Summary

In sum, disentangling how working and long-term memory interact is difficult because they so readily collaborate with one another. Yet, carefully characterizing this interaction is key for understanding how memory functions in realistic settings. Here, we introduced a controlled opportunity for long-term memory to assist working memory by combining abstracted stimuli with explicit repetitions of memory arrays. Participants were successful at rapidly recruiting long-term memory to assist performance – they reached perfect performance for supra-capacity arrays after only a few repetitions. This rapid learning for even highly similar and arbitrary arrays is illustrative of how difficult it is to get a "process pure" measure of working memory in the absence of any support from long-term memory. Looking forward, we think that leaning in to the natural collaboration of working- and long-term memory will be key for furthering our understanding of each memory system, and there is much future work to be done to build this understanding (e.g., probing neural measures, strategy manipulations, and task constraints).

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.3758/s13421-024-01566-z>.

Acknowledgements We thank Rebecca Kuang for assistance with data collection.

Author Contributions KA: Conceptualization, methodology, validation, formal analysis, investigation, writing - original draft, writing - review and editing, visualization. CZ: Validation, writing - review and editing. EV: Funding acquisition, writing - review and editing.

Funding NSF SBE Postdoctoral Fellowship 2104630 to K.A. and grants ONR-N000142212123 and R01-MH087214 to E.V.

Data availability Raw data and code to reproduce analyses will be made available on the Open Science Framework upon publication (<https://osf.io/dw7cp/>).

¹¹ Although increased encoding time has not been shown to improve visual working memory performance, this contrasts with the verbal literature where encoding time has been shown to improve recall (e.g., Waugh, 1967).

Declarations

Conflicts of interest None.

References

- Adam, K. C. S., Mance, I., Fukuda, K., & Vogel, E. K. (2015). The Contribution of Attentional Lapses to Individual Differences in Visual Working Memory Capacity. *Journal of Cognitive Neuroscience*, 27(8), 1601–1616. https://doi.org/10.1162/jocn_a_00811
- Anderson, J. R., & Matessa, M. (1997). A production system theory of serial memory. *Psychological Review*, 104(4), 728–748. <https://doi.org/10.1037/0033-295X.104.4.728>
- Asp, I. E., Störmer, V. S., & Brady, T. F. (2021). Greater Visual Working Memory Capacity for Visually Matched Stimuli When They Are Perceived as Meaningful. *Journal of Cognitive Neuroscience*, 33(5), 902–918. https://doi.org/10.1162/jocn_a_01693
- Baddeley, A. D., & Hitch, G. (1974). Working Memory. In *Psychology of Learning and Motivation* (Vol. 8, pp. 47–89). Elsevier. <http://linkinghub.elsevier.com/retrieve/pii/S0079742108604521>. Accessed 24 Aug 2016
- Baddeley, A. D., & Warrington, E. K. (1970). Amnesia and the distinction between long- and short-term memory. *Journal of Verbal Learning and Verbal Behavior*, 9(2), 176–189. [https://doi.org/10.1016/S0022-5371\(70\)80048-2](https://doi.org/10.1016/S0022-5371(70)80048-2)
- Banks, W. P. (1970). Signal detection theory and human memory. *Psychological Bulletin*, 74(2), 81–99. <https://doi.org/10.1037/h0029531>
- Beaudoin, M., & Desrichard, O. (2011). Are memory self-efficacy and memory performance related? A meta-analysis. *Psychological Bulletin*, 137(2), 211–241. <https://doi.org/10.1037/a0022106>
- Beck, M. R., & van Lamsweerde, A. E. (2011). Accessing long-term memory representations during visual change detection. *Memory & Cognition*, 39(3), 433–446. <https://doi.org/10.3758/s13421-010-0033-4>
- Bousfield, W. A., & Sedgewick, C. H. W. (1944). An Analysis of Sequences of Restricted Associative Responses. *The Journal of General Psychology*, 30(2), 149–165. <https://doi.org/10.1080/00221309.1944.10544467>
- Bower, G. H. (1972). Perceptual groups as coding units in immediate memory. *Psychonomic Science*, 27(4), 217–219. <https://doi.org/10.3758/BF03328942>
- Brady, T. F., Konkle, T., Alvarez, G. A., & Oliva, A. (2008). Visual long-term memory has a massive storage capacity for object details. *Proceedings of the National Academy of Sciences of the United States of America*, 105(38), 14325–14329. <https://doi.org/10.1073/pnas.0803390105>
- Brady, T. F., & Störmer, V. S. (2022). The role of meaning in visual working memory: Real-world objects, but not simple features, benefit from deeper processing. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 48(7), 942–958. <https://doi.org/10.1037/xlm0001014>
- Brady, T. F., Störmer, V. S., & Alvarez, G. A. (2016). Working memory is not fixed-capacity: More active storage capacity for real-world objects than for simple stimuli. *Proceedings of the National Academy of Sciences of the United States of America*, 113(27), 7459–7464. <https://doi.org/10.1073/pnas.1520027113>
- Brainard, D. H. (1997). The Psychophysics Toolbox. *Spatial Vision*, 10(4), 433–436. <https://doi.org/10.1163/156856897X00357>
- Broadbent, D. E. (1975). The Magic Number Seven After Fifteen Years. In A. Kennedy & A. Wilkes (Eds.), *Studies In Long Term Memory*. John Wiley & Sons.
- Browman, C. P., & O’Connell, D. C. (1976). Sequential phonological effects in recitation times. *Bulletin of the Psychonomic Society*, 8(1), 37–39. <https://doi.org/10.3758/BF03337067>
- Carlisle, N. B., Arita, J. T., Pardo, D., & Woodman, G. F. (2011). Attentional Templates in Visual Working Memory. *Journal of Neuroscience*, 31(25), 9315–9322. <https://doi.org/10.1523/JNEUROSCI.1097-11.2011>
- Carlson, R. A., Wenger, J. L., & Sullivan, M. A. (1993). Coordinating information from perception and working memory. *Journal of Experimental Psychology: Human Perception and Performance*, 19(3), 531–548. <https://doi.org/10.1037//0096-1523.19.3.531>
- Chase, W. G., & Ericsson, K. A. (1982). Skill and Working Memory. In G. H. Bower (Ed.), *Psychology of Learning and Motivation* (Vol. 16, pp. 1–58). Academic Press. [https://doi.org/10.1016/S0079-7421\(08\)60546-0](https://doi.org/10.1016/S0079-7421(08)60546-0)
- Chase, W. G., & Simon, H. A. (1973). Perception in chess. *Cognitive Psychology*, 4(1), 55–81. [https://doi.org/10.1016/0010-0285\(73\)90004-2](https://doi.org/10.1016/0010-0285(73)90004-2)
- Christophel, T. B., Klink, P. C., Spitzer, B., Roelfsema, P. R., & Haynes, J.-D. (2017). The Distributed Nature of Working Memory. *Trends in Cognitive Sciences*, 21(2), 111–124. <https://doi.org/10.1016/j.tics.2016.12.007>
- Cowan, N. (2001). The magical number 4 in short-term memory: a reconsideration of mental storage capacity. *The Behavioral and Brain Sciences*, 24(1), 87–114; discussion 114–185. <https://doi.org/10.1017/S0140525X01003922>
- Craik, F. I. M., & Lockhart, R. S. (1972). Levels of processing: A framework for memory research. *Journal of Verbal Learning and Verbal Behavior*, 11(6), 671–684. [https://doi.org/10.1016/S0022-5371\(72\)80001-X](https://doi.org/10.1016/S0022-5371(72)80001-X)
- Dunlosky, J., & Salthouse, T. A. (1996). A decomposition of age-related differences in multitrial free recall. *Neuropsychology, Development, and Cognition. Section B, Aging, Neuropsychology and Cognition*, 3(1), 2–14. <https://doi.org/10.1080/13825589608256608>
- Ebbinghaus, H. (1885). *Über das Gedächtnis: Untersuchungen zur Experimentellen Psychologie*. Verlag von Duncker & Humblot.
- Ebbinghaus, H. (1913). *Memory: A contribution to experimental psychology* (H. A. Ruger & C. E. Bussenius, trans.). Teachers College, Columbia University.
- Endress, A. D., & Potter, M. C. (2014). Large capacity temporary visual memory. *Journal of Experimental Psychology: General*, 143(2), 548–565. <https://doi.org/10.1037/a0033934>
- Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, 39(2), 175–191. <https://doi.org/10.3758/BF03193146>
- Forsberg, A., Johnson, W., & Logie, R. H. (2020). Cognitive aging and verbal labeling in continuous visual memory. *Memory & Cognition*, 48(7), 1196–1213. <https://doi.org/10.3758/s13421-020-01043-3>
- Fukuda, K., & Vogel, E. K. (2019). Visual short-term memory capacity predicts the “bandwidth” of visual long-term memory encoding. *Memory & Cognition*, 47(8), 1481–1497. <https://doi.org/10.3758/s13421-019-00954-0>
- Graesser, A., & Mandler, G. (1978). Limited processing capacity constrains the storage of unrelated sets of words and retrieval from natural categories. *Journal of Experimental Psychology: Human Learning and Memory*, 4(1), 86–100. <https://doi.org/10.1037/0278-7393.4.1.86>
- Hebb, D. (1961). Distinctive features of learning in the higher animal. In J. F. Delafresnaye (Ed.), *Brain mechanisms and learning* (pp. 37–46). Oxford University Press. <https://www.semanticscholar.org/paper/5705ba6782a7016182093e6d3f9ff2425e581659>
- Heinen, K. T. H., Kenemans, J. L., & van der Stigchel, S. (2022). Recruitment of a long-term memory supporting neural network

- during repeated maintenance of a multi-item abstract visual image in working memory. *Scientific Reports*, 12(1), 575. <https://doi.org/10.1038/s41598-021-04384-4>
- Hertzog, C., Dixon, R. A., & Hulstsch, D. F. (1990). Relationships between metamemory, memory predictions, and memory task performance in adults. *Psychology and Aging*, 5(2), 215–227. <https://doi.org/10.1037/0882-7974.5.2.215>
- Hertzog, C., Saylor, L. L., Fleece, A. M., & Dixon, R. A. (1994). Metamemory and aging: Relations between predicted, actual and perceived memory task performance. *Neuropsychology, Development, and Cognition. Section B, Aging, Neuropsychology and Cognition*, 1(3), 203–237. <https://doi.org/10.1080/13825589408256577>
- Huang, L. (2010). Visual working memory is better characterized as a distributed resource rather than discrete slots. *Journal of Vision*, 10(14), 8–8. <https://doi.org/10.1167/10.14.8>
- Huang, L., & Awh, E. (2018). Chunking in working memory via content-free labels. *Scientific Reports*, 8(1), 23. <https://doi.org/10.1038/s41598-017-18157-5>
- Jackson, M. C., & Raymond, J. E. (2008). Familiarity enhances visual working memory for faces. *Journal of Experimental Psychology. Human Perception and Performance*, 34(3), 556–568. <https://doi.org/10.1037/0096-1523.34.3.556>
- Jenson, A., & Squire, L. R. (2011). Working memory, long-term memory, and medial temporal lobe function. *Learning & Memory*, 19(1), 15–25. <https://doi.org/10.1101/lm.024018.111>
- Jolly, E. (2018). Pym4: Connecting R and python for linear mixed modeling. *Journal of Open Source Software*, 3(31), 862. <https://doi.org/10.21105/joss.00862>
- Kim, H.-Y. (2016). Statistical notes for clinical researchers: Sample size calculation 3. Comparison of several means using one-way ANOVA. *Restorative Dentistry & Endodontics*, 41(3), 231–234. <https://doi.org/10.5395/rde.2016.41.3.231>
- Kleiner, M., Brainard, D., & Pelli, D. (2007). *What's new in Psychtoolbox-3?* European Conference on Visual Perception (ECVP). <https://pdfs.semanticscholar.org/04d4/7572cec08b7a582a9366e5ac61dcfd633f2a.pdf>
- Kluyver, T., Ragan-Kelley, B., Pérez, F., Granger, B., Bussonnier, M., Frederic, J., Kelley, K., Hamrick, J., Grout, J., Corlay, S., Ivanov, P., Avila, D., Abdalla, S., Willing, C., & Jupyter Development Team. (2016). Jupyter Notebooks – a publishing format for reproducible computational workflows. In *Positioning and Power in Academic Publishing: Players, Agents and Agendas* (pp. 87–90). IOS Press. <https://doi.org/10.3233/978-1-61499-649-1-87>
- Logie, R. H., Brockmole, J. R., & Vandenbroucke, A. R. E. (2009). Bound feature combinations in visual short-term memory are fragile but influence long-term learning. *Visual Cognition*, 17(1–2), 160–179. <https://doi.org/10.1080/13506280802228411>
- Lovelace, E. A., & Snodgrass, R. D. (1971). Decision times for alphabetic order of letter pairs. *Journal of Experimental Psychology*, 88(2), 258–264. <https://doi.org/10.1037/h0030922>
- Lovelace, E. A., & Spence, W. A. (1972). Reaction times for naming successive letters of the alphabet. *Journal of Experimental Psychology*, 94(2), 231–233. <https://doi.org/10.1037/h0032763>
- Luck, S. J., & Vogel, E. K. (1997). The capacity of visual working memory for features and conjunctions. *Nature*, 390(6657), 279–281. <https://doi.org/10.1038/36846>
- Mandler, J. M., & Ritchey, G. H. (1977). Long-term memory for pictures. *Journal of Experimental Psychology. Human Learning and Memory*, 3(4), 386–396. <https://doi.org/10.1037/0278-7393.3.4.386>
- McKinney, W. (2010). Data Structures for Statistical Computing in Python. *Proceedings of the 9th Python in Science Conference*. Python in Science Conference, Austin, Texas. <https://doi.org/10.25080/majora-92bf1922-00a>
- McLean, R. S., & Gregg, L. W. (1967). Effects of induced chunking on temporal aspects of serial recitation. *Journal of Experimental Psychology*, 74(4), 455–459. <https://doi.org/10.1037/h0024785>
- Miller, A. L., & Unsworth, N. (2018). Individual differences in working memory capacity and search efficiency. *Memory & Cognition*, 46(7), 1149–1163. <https://doi.org/10.3758/s13421-018-0827-3>
- Milner, B., & Penfield, W. (1955). The effect of hippocampal lesions on recent memory. *Transactions of the American Neurological Association, 80th Meeting*, 42–48. <https://www.ncbi.nlm.nih.gov/pubmed/13311995>
- Murdock, B. B., & Okada, R. (1970). Interresponse times in single-trial free recall. *Journal of Experimental Psychology*, 86(2), 263–267. <https://doi.org/10.1037/h0029993>
- Musfeld, P., Souza, A. S., & Oberauer, K. (2023). Repetition learning is neither a continuous nor an implicit process. *Proceedings of the National Academy of Sciences of the United States of America*, 120(16), e2218042120. <https://doi.org/10.1073/pnas.2218042120>
- Musfeld, P., Souza, A. S., & Oberauer, K. (2023). Testing expectations and retrieval practice modulate repetition learning of visuospatial arrays. *Journal of Experimental Psychology. Learning, Memory, and Cognition*. <https://doi.org/10.1037/xlm0001298>
- Ngiam, W. X. Q., Brissenden, J. A., & Awh, E. (2019a). “Memory compression” effects in visual working memory are contingent on explicit long-term memory. *Journal of Experimental Psychology. General*, 148(8), 1373–1385. <https://doi.org/10.1037/xge0000649>
- Ngiam, W. X. Q., Khaw, K. L. C., Holcombe, A. O., & Goodbourn, P. T. (2019b). Visual working memory for letters varies with familiarity but not complexity. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, 45(10), 1761–1775. <https://doi.org/10.1037/xlm0000682>
- Olson, I. R., & Jiang, Y. (2004). Visual short-term memory is not improved by training. *Memory & Cognition*, 32(8), 1326–1332. <https://doi.org/10.3758/BF03206323>
- Overkott, C., & Souza, A. S. (2021). Verbal descriptions improve visual working memory but have limited impact on visual long-term memory. *Journal of Experimental Psychology. General*. <https://doi.org/10.1037/xge0001084>
- Page, M. P. A., Cumming, N., Norris, D., McNeil, A. M., & Hitch, G. J. (2013). Repetition-spacing and item-overlap effects in the Hebb repetition task. *Journal of Memory and Language*, 69(4), 506–526. <https://doi.org/10.1016/j.jml.2013.07.001>
- Quirk, C., Adam, K. C. S., & Vogel, E. K. (2020). No Evidence for an Object Working Memory Capacity Benefit with Extended Viewing Time. *ENEURO*, 7(5), ENEURO.0150–20.2020. <https://doi.org/10.1523/ENEURO.0150-20.2020>
- Reder, L. M., Victoria, L. W., Manelis, A., Oates, J. M., Dutcher, J. M., Bates, J. T., Cook, S., Aizenstein, H. J., Quinlan, J., & Gyulai, F. (2013). Why it's easier to remember seeing a face we already know than one we don't: preexisting memory representations facilitate memory formation. *Psychological Science*, 24(3), 363–372. <https://doi.org/10.1177/0956797612457396>
- Reinhart, R. M. G., & Woodman, G. F. (2014). High stakes trigger the use of multiple memories to enhance the control of attention. *Cerebral Cortex*, 24(8), 2022–2035. <https://doi.org/10.1093/cercor/bht057>
- Reitman, J. S. (1976). Skilled perception in Go: Deducing memory structures from inter-response times. *Cognitive Psychology*, 8(3), 336–356. [https://doi.org/10.1016/0010-0285\(76\)90011-6](https://doi.org/10.1016/0010-0285(76)90011-6)
- Rohatgi, A. (2022). *WebPlotDigitizer*. <https://automeris.io/WebPlotDigitizer>
- Rohrer, D. (1996). On the relative and absolute strength of a memory trace. *Memory & Cognition*, 24(2), 188–201. <https://doi.org/10.3758/bf03200880>
- 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>

- Scoville, W. B., & Milner, B. (1957). LOSS OF RECENT MEMORY AFTER BILATERAL HIPPOCAMPAL LESIONS. *Journal of Neurology, Neurosurgery, and Psychiatry*, 20(1), 11–21. <https://doi.org/10.1136/jnmp.20.1.11>
- Sense, F., Morey, C. C., Prince, M., Heathcote, A., & Morey, R. D. (2017). Opportunity for verbalization does not improve visual change detection performance: A state-trace analysis. *Behavior Research Methods*, 49(3), 853–862. <https://doi.org/10.3758/s13428-016-0741-1>
- Serences, J. T., Ester, E. F., Vogel, E. K., & Awh, E. (2009). Stimulus-Specific Delay Activity in Human Primary Visual Cortex. *Psychological Science*, 20(2), 207–214. <https://doi.org/10.1111/j.1467-9280.2009.02276.x>
- Souza, A. S., & Oberauer, K. (2022). Promoting visual long-term memories: When do we learn from repetitions of visuospatial arrays? *Journal of Experimental Psychology. General*, 151(12), 3114–3133. <https://doi.org/10.1037/xge0001236>
- Souza, A. S., Overkott, C., & Matyja, M. (2021). Categorical distinctiveness constrains the labeling benefit in visual working memory. *Journal of Memory and Language*. <https://www.sciencedirect.com/science/article/pii/S0749596X21000255>
- Squire, L. R., & Zola, S. M. (1996). Structure and function of declarative and nondeclarative memory systems. *Proceedings of the National Academy of Sciences*, 93(24), 13515–13522. <https://doi.org/10.1073/pnas.93.24.13515>
- Standing, L. (1973). Learning 10000 pictures. *The Quarterly Journal of Experimental Psychology*, 25(2), 207–222. <https://doi.org/10.1080/14640747308400340>
- Standing, L., Conezio, J., & Haber, R. N. (1970). Perception and memory for pictures: Single-trial learning of 2500 visual stimuli. *Psychonomic Science*, 19(2), 73–74. <https://doi.org/10.3758/BF03337426>
- Sternberg, R. J., & Tulving, E. (1977). The measurement of subjective organization in free recall. *Psychological Bulletin*, 84(3), 539–556. <https://doi.org/10.1037/0033-2909.84.3.539>
- Sukegawa, M., Ueda, Y., & Saito, S. (2019). The effects of Hebb repetition learning and temporal grouping in immediate serial recall of spatial location. *Memory & Cognition*, 47(4), 643–657. <https://doi.org/10.3758/s13421-019-00921-9>
- Tsubomi, H., Fukuda, K., Watanabe, K., & Vogel, E. K. (2013). Neural Limits to Representing Objects Still within View. *Journal of Neuroscience*, 33(19), 8257–8263. <https://doi.org/10.1523/JNEUROSCI.5348-12.2013>
- Tulving, E. (1962). Subjective organization in free recall of “unrelated” words. *Psychological Review*, 69, 344–354. <https://doi.org/10.1037/h0043150>
- Tulving, E. (1966). Subjective organization and effects of repetition in multi-trial free-recall learning. *Journal of Verbal Learning and Verbal Behavior*, 5(2), 193–197. [://doi.org/https://doi.org/10.1016/S0022-5371\(66\)80016-6](https://doi.org/10.1016/S0022-5371(66)80016-6)
- Unsworth, N., & Engle, R. W. (2005). Individual differences in working memory capacity and learning: evidence from the serial reaction time task. *Memory & Cognition*, 33(2), 213–220. <https://doi.org/10.3758/bf03195310>
- Vallat, R. (2018). Pingouin: statistics in Python. *Journal of Open Source Software*, 3(31), 1026. <https://doi.org/10.21105/joss.01026>
- Waskom, M. (2021). seaborn: statistical data visualization. *Journal of Open Source Software*, 6(60), 3021. <https://doi.org/10.21105/joss.03021>
- Waugh, N. C. (1967). Presentation time and free recall. *Journal of Experimental Psychology*, 73(1), 39–44. <https://doi.org/10.1037/h0024125>
- West, R. L., Dennehy-Basile, D., & Norris, M. P. (1996). Memory self-evaluation: The effects of age and experience. *Neuropsychology, Development, and Cognition. Section B, Aging, Neuropsychology and Cognition*, 3(1), 67–83. <https://doi.org/10.1080/13825589608256613>
- Wixted, J. T., & Rohrer, D. (1994). Analyzing the dynamics of free recall: An integrative review of the empirical literature. *Psychonomic Bulletin & Review*, 1(1), 89–106. <https://doi.org/10.3758/BF03200763>
- Wolfe, J. M., Wick, F. A., Mishra, M., DeGutis, J., & Lyu, W. (2023). Spatial and temporal massive memory in humans. *Current Biology: CB*, 33(2), 405–410.e4. <https://doi.org/10.1016/j.cub.2022.12.040>
- Xie, W., & Zhang, W. (2017). Familiarity increases the number of remembered Pokémon in visual short-term memory. *Memory & Cognition*, 45(4), 677–689. <https://doi.org/10.3758/s13421-016-0679-7>
- Xie, W., & Zhang, W. (2022). Pre-existing long-term memory facilitates the formation of visual short-term memory. In *Visual Memory* (1st Edition, pp. 84–104). Routledge. <https://doi.org/10.4324/9781003158134-6>

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.