

Retrieval practice enhances the accessibility but not the quality of memory

David W. Sutterer^{1,2} · Edward Awh^{1,2,3}

© Psychonomic Society, Inc. 2015

Abstract Numerous studies have demonstrated that retrieval from long-term memory (LTM) can enhance subsequent memory performance, a phenomenon labeled the retrieval practice effect. However, the almost exclusive reliance on categorical stimuli in this literature leaves open a basic question about the nature of this improvement in memory performance. It has not yet been determined whether retrieval practice improves the probability of successful memory retrieval or the quality of the retrieved representation. To answer this question, we conducted three experiments using a mixture modeling approach (Zhang & Luck, 2008) that provides a measure of both the probability of recall and the quality of the recalled memories. Subjects attempted to memorize the color of 400 unique shapes. After every 10 images were presented, subjects either recalled the last 10 colors (the retrieval practice condition) by clicking on a color wheel with each shape as a retrieval cue or they participated in a control condition that involved no further presentations (Experiment 1) or restudy of the 10 shape/color associations (Experiments 2 and 3). Performance

in a subsequent delayed recall test revealed a robust retrieval practice effect. Subjects recalled a significantly higher proportion of items that they had previously retrieved relative to items that were untested or that they had restudied. Interestingly, retrieval practice did not elicit any improvement in the precision of the retrieved memories. The same empirical pattern also was observed following delays of greater than 24 hours. Thus, retrieval practice increases the probability of successful memory retrieval but does not improve memory quality.

Keywords Cued recall · Memory · Mnemonic precision · Testing effect

Introduction

Numerous studies have demonstrated that retrieval from long-term memory (LTM) can enhance subsequent memory performance, a phenomenon labeled the retrieval practice effect (Carrier & Pashler, 1992). The benefits of retrieval practice have been observed with a wide variety of memoranda (Roediger & Karpicke, 2006), including word pairs (Pyc & Rawson, 2009), pictures (Wheeler & Roediger, 1992), and spatial positions (Carpenter & Pashler 2007; Rohrer, Taylor, and Sholar, 2010; Carpenter & Kelly, 2012).

Varying explanations have been offered for how retrieval practice enhances memory performance. Some have focused on increased elaborative retrieval during testing (Carpenter, 2009), whereas others have emphasized the narrowing of the retrieval search space via helpful contextual associations (Lehman, Smith, and Karpicke, 2014). One common assumption of these accounts is that retrieval practice enhances the probability of access to a memory rather than the quality of the memory. This focus on accessibility over fidelity may be

Electronic supplementary material The online version of this article (doi:10.3758/s13423-015-0937-x) contains supplementary material, which is available to authorized users.

✉ David W. Sutterer
sutterer@uchicago.edu

✉ Edward Awh
awh@uchicago.edu

¹ Department of Psychology, The University of Chicago, Chicago, IL, USA

² The Institute for Mind and Biology, The University of Chicago, 940 E. 57th St., Chicago, IL 60637, USA

³ The Grossman Institute for Neuroscience, The University of Chicago, 940 E. 57th St., Chicago, IL 60637, USA

attributable in part to the fact that past studies have typically used discrete word or picture stimuli (and all-or-none measures of accuracy) that do not allow clear measurements of memory fidelity. That said, some past findings may be consistent with a putative effect of retrieval practice on memory quality. For example, Chan and McDermott (2007) found that retrieval practice improved participants' ability to avoid semantically similar lures during a recognition test and improved source memory. Likewise, Szpunar, McDermott, and Roediger (2008) found that testing improves list discrimination. However, while each of these findings could reflect a more precise memory (e.g., of specific semantic content, or of the temporal context associated with an item), the binary nature of the responses in these studies also allows for an interpretation based on retrieval probability.

An approach that may provide more traction for understanding the effect of retrieval practice on the quality of item-specific memory is to allow participants to report remembered information along a continuous response space. For example, Carpenter and Kelly (2012) used a continuous response space in a task where subjects recalled the precise positions of different objects. Retrieval practice resulted in a decrease in the average response error for retrieved locations relative to restudied locations. However, although a change in memory quality provides an intuitive explanation of these findings, a reduced guessing rate in the retrieval practice condition also would yield lower average response errors. Thus, the goal of the present work was to examine the retrieval practice effect using an analytic approach that can estimate both the probability of retrieval and the quality of the retrieved representations.

We measured performance in a shape/color recall task in which the possible colors were drawn from a continuous 360-degree space, and we used a mixture-modeling approach (Zhang & Luck, 2008) that provided separate measures of the probability of recall and the quality of the retrieved memories. This analytic approach has been widely applied to the field of working memory (see Luck & Vogel, 2013 for review), and has recently been applied to the study of LTM (Brady et al., 2013). To anticipate our conclusions, retrieval practice elicited robust improvements in the probability of memory access, but absolutely no improvement in the fidelity of the retrieved memories.

Experiment 1: Test versus no test

Method

Participants

Twenty-two undergraduates at the University of Oregon completed the experiment for course credit. All participants gave

informed consent according to procedures approved by the University of Oregon institutional review board.

Apparatus

Stimuli were generated in MATLAB using Psychophysics Toolbox extension (Brainard, 1997; Pelli, 1997) and were presented on a 17-in. flat CRT computer screen (60-HZ refresh rate). The viewing distance was ~80 cm. Stimuli were $9.2^\circ \times 9.2^\circ$ of visual angle.

Stimuli

Four hundred nameable pictures (e.g., animals, plants, shapes, countries, U.S. states, and symbols) were obtained via a web search for royalty free clip art. One of 360 continuous colors was assigned to each image, with different color/shape sets for each subject.

Task and procedure

The 400 stimuli were presented in two successive runs, each containing 200 distinct shape/color associations. Each run was comprised of two parts: a learning period and a delayed-retrieval period. During the learning period, images were presented serially in blocks of 10 items, followed either by retrieval practice, during which all 10 colors were recalled or by the start of the next block of 10 items (Fig. 1); thus, subjects did not know during encoding whether or not they would be immediately tested. Images were tested in a random order without feedback.

After viewing all 200 images with retrieval practice for half of the items in the run (~20-30 minutes), subjects were asked to recall the color of each image by clicking on a color wheel that represented all of the presented colors. Images were tested in a random order relative to their initial presentation. Participants received feedback consisting of the presentation of the shape filled with the correct color and a number denoting the magnitude of the error.

During recall, a white shape cue was displayed for 1 second before the cursor and color wheel appeared (Fig. 1B). During response selection, the color of the shape cue shifted continuously to match the hue that was indicated by the mouse cursor on the color wheel. Participants indicated their color choice by clicking the mouse. Responses were unspeeded and accuracy was given highest priority; subjects were instructed to choose a response even if they felt they were guessing. When they thought they were guessing, they were instructed to click with the right mouse button rather than the left. The color wheel was randomly rotated across trials (so that position information was irrelevant to the color response). Following completion of the first run of 200 images, the remaining 200 images were presented and tested using the same procedure (i.e., a

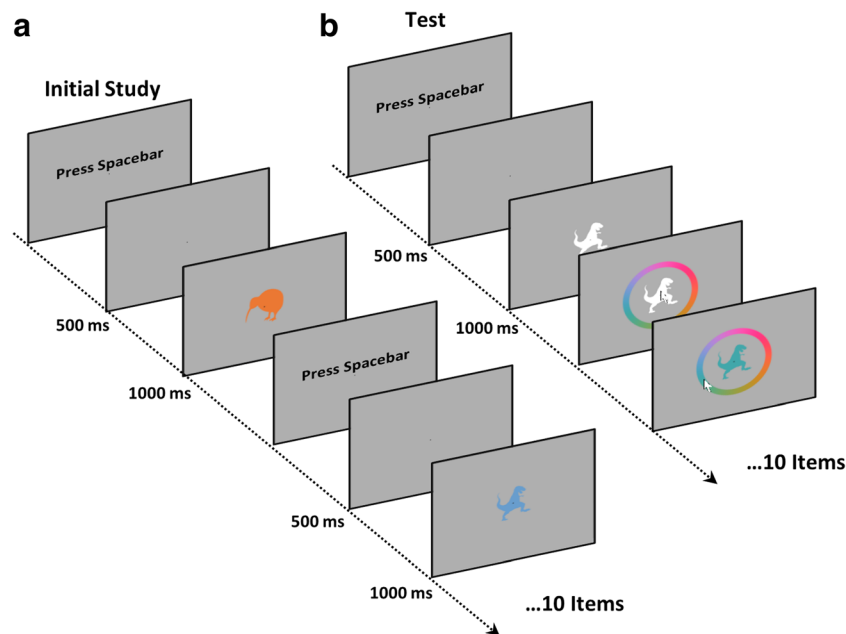


Fig. 1 Task diagram for Experiment 1. **(A)** Timing of stimuli for initial study opportunity. Participants studied 10 stimuli and then were either tested on those ten images or not tested. Subjects studied 400 images total. **(B)** Timing of each item at initial test and at final test

learning period and delayed-retrieval period) with 200 new images. One image was presented twice during the learning period of run one and was dropped from the delayed analyses.

Data analysis

Response error was measured as the number of degrees between the presented color and the reported color. Errors ranged from 0° (perfect response) to $\pm 180^\circ$ (a maximally imprecise response). Responses were centered on 0° but spanned the entire range of responses (for example, see Fig. 2A). These error histograms are well described as a mixture of two distributions that reflect guesses and correct responses (Zhang & Luck, 2008). On some trials, subjects do not remember the color associated with the shape cue and guess randomly with respect to the target color. This results in a uniform distribution of responses with respect to the target color. On other trials, participants remember the color of the shape cue and provide responses centered on the correct color value but with some degree of error. This distribution is well described by a von Mises distribution (the circular analogue of a Gaussian distribution because the tested color space was circular) centered on the correct response. To obtain an estimate of these two distributions, response errors were fit using Markov Chain Monte Carlo (MCMC) as employed by the “memfit” function of Memtoolbox (Suchow, Brady, Fougner, and Alvarez, 2013). MCMC repeatedly samples parameter values in proportion to how well they describe the data and the prior (in this case an uninformative

Jeffreys prior) to obtain a Maximum a Posteriori (MAP) estimate of three parameters: P_{mem} is the probability that subjects could retrieve nonzero target information, operationalized as the inverse of the height of the uniform distribution (i.e., $1 - \text{proportion of guesses}$). SD is the standard deviation of the von Mises distribution (with larger values reflecting reduced precision). Mu (μ), the mean of the von Mises distribution, reflects systematic bias in the error distribution (preferred clockwise or anti-clockwise responses on the color wheel).¹

These parameters are calculated using the distribution of all responses, which is a mixture of responses not guided by memory (guesses) and responses guided by memory. Thus, we can determine the proportion of remembered items and the precision of responses guided by memory, but it is not possible to determine if any individual response was guided by memory.

Results

Aggregate data

All participants’ responses were combined into an aggregate error histogram (Fig. 2A) and fit using the “memfit” function of Memtoolbox (Suchow et al. 2013) to obtain parameter

¹ A coding error resulted in a 1° rotation of the color wheel relative to the correct position, and we have detailed the negligible results of the effect on μ in Supplementary Table 1. Results of μ will not be discussed further.

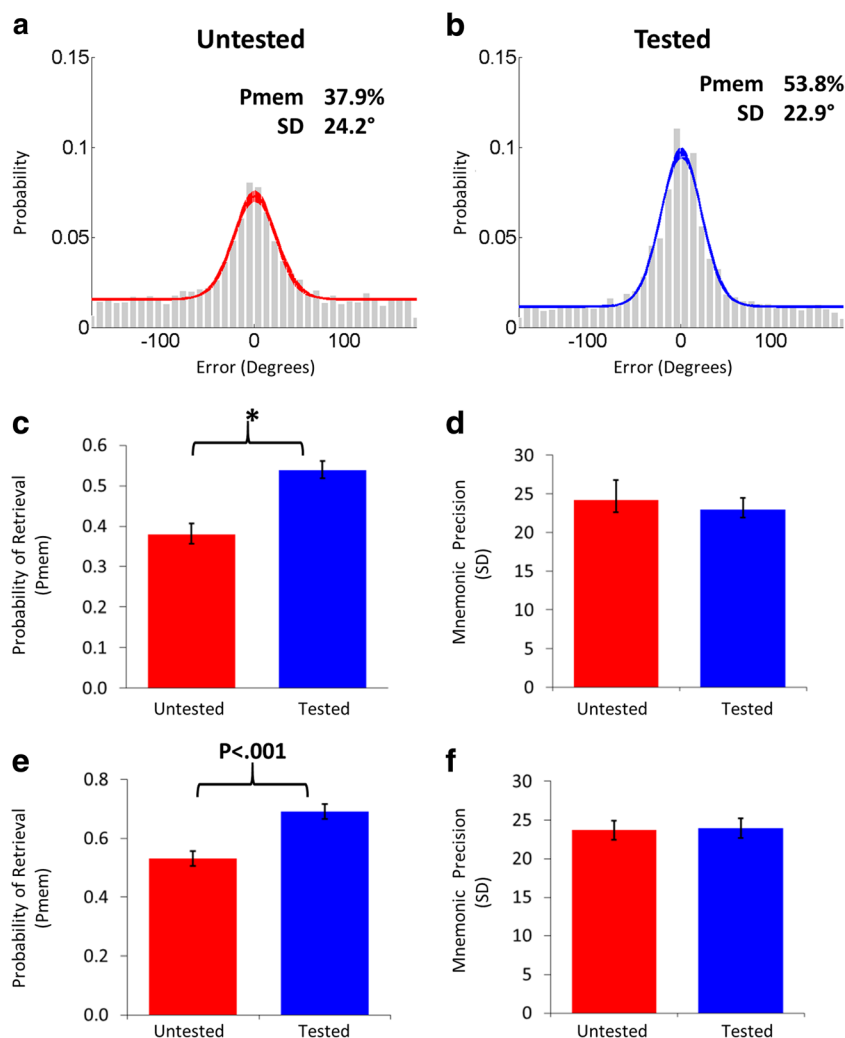


Fig. 2 Delayed test results from Experiment 1. **(A)** Aggregate fit of delayed response errors from all subjects for initially untested items. **(B)** Aggregate fit of delayed response errors from all subjects for initially retrieved items. **(C+D)** Aggregate parameter estimates of probability of retrieval (P_{mem}) and mnemonic precision (SD) at delayed test*. Error bars represent Bayesian credibility intervals of the fits. **(E+F)** Average of

individual parameter estimates of P_{mem} and SD for subjects who successfully retrieved $>40\%$ of the items at the delayed test. Error bars represent 95% within-subjects confidence intervals that were calculated by normalizing to remove within subject variance (Loftus & Masson, 1994). *Denotes nonoverlapping credibility intervals

estimates and 95% credibility intervals (CrI); there is a 95% chance that the true value of the parameter for the sample lies between the credibility intervals. We will refer to parameters with overlapping credibility intervals as “not significantly different” and parameters with nonoverlapping credibility intervals as “significantly different.” Unlike confidence intervals, Bayesian credibility intervals are not necessarily symmetrical.

The mixture modeling analysis revealed that 70.7% (CrI: -1.7% , $+2.0\%$) of the items were recalled during the initial test. SD—our operational definition of mnemonic precision—was 21.4° (CrI: -0.8° , $+1.1^\circ$). At delayed test, subjects recalled significantly more items that they had previously retrieved (53.8%, CrI: -1.9% , $+2.3\%$) than items that that were previously untested (37.9%, CrI: -2.2% , $+2.8\%$; Fig. 2). Mnemonic precision was not significantly different between tested

(22.9°, CrI: -1.0° , $+1.5^\circ$) and untested (24.2°, CrI: -1.6° , $+2.6^\circ$) items.

Simulations

We were interested in examining the data at the individual subject level, but simulations showed that there would be consistent biases in the precision estimates if the probability of retrieval was too low. We determined this by generating artificial data that presumed varying P_{mem} values and SD values equal to those observed in our aggregate data (20°). Parameter estimates were obtained from these artificial datasets by sampling 100 times from each dataset and then fitting each sample with a mixture model. These simulations revealed that SD is systematically overestimated when the proportion of

successfully retrieved items was less than 40% (Fig. 3A). By contrast, the P_{mem} parameter is relatively accurate even when probability of retrieval is low. Thus, to avoid misleading estimates of SD, we compared individual parameter estimates of precision only for subjects who successfully retrieved at least 40% of the items in both the tested and untested conditions. Further simulations confirmed that estimates of the SD parameter would not be affected by high guess rates in the aggregate data because of the large number of trials run across all subjects (>4,000 trials per condition). Thus, in the aggregate analysis, accurate P_{mem} and SD estimates could be obtained even when probability of retrieval was low (Fig. 3B).

Individual parameter comparisons (Delayed Test)

Analysis of the subset of subjects who successfully retrieved 40% or more items in both conditions ($n = 12$) also showed higher P_{mem} for retrieved items ($M = 69.2\%$, $SD = 13.4\%$) compared with untested items ($M = 53.2\%$, $SD = 9.9\%$, $t(11) = -6.03$; $p < 0.001$). Also in line with the aggregate data, subjects did not exhibit superior mnemonic precision for items that they had previously retrieved ($M = 24.0$, $SD = 5.3$) compared with items that were not retrieved ($M = 23.7$, $SD = 5.7$, $t(11) = -0.22$ $p = 0.83$).

Discussion

Experiment 1 suggests that retrieval practice increases the probability that an item can be retrieved in the future but does

not improve the precision of that memory. In Experiment 2, we equated the number of times that participants saw and responded to each item by comparing the retrieval practice condition with a restudy condition (Carrier & Pashler, 1992).

Experiment 2

Method

Participants

Twenty-eight students from the University of Oregon participated in Experiment 2 for course credit or monetary compensation. Six participants were excluded: two did not complete the session, one was excluded during the session for not following instructions, and three participants who completed the session were excluded for responding randomly on restudy trials. Twenty-two participants were included in the analysis of Experiment 2. Six subjects who did not complete all trials in the time allotted were included in the experiment, because they had completed the session and followed instructions.

Task

The task in Experiment 2 was the same as Experiment 1, except that during the learning period subjects either completed a memory test or were given a chance to restudy the items after every 10 images (Fig. 4A). During restudy, the shape cue

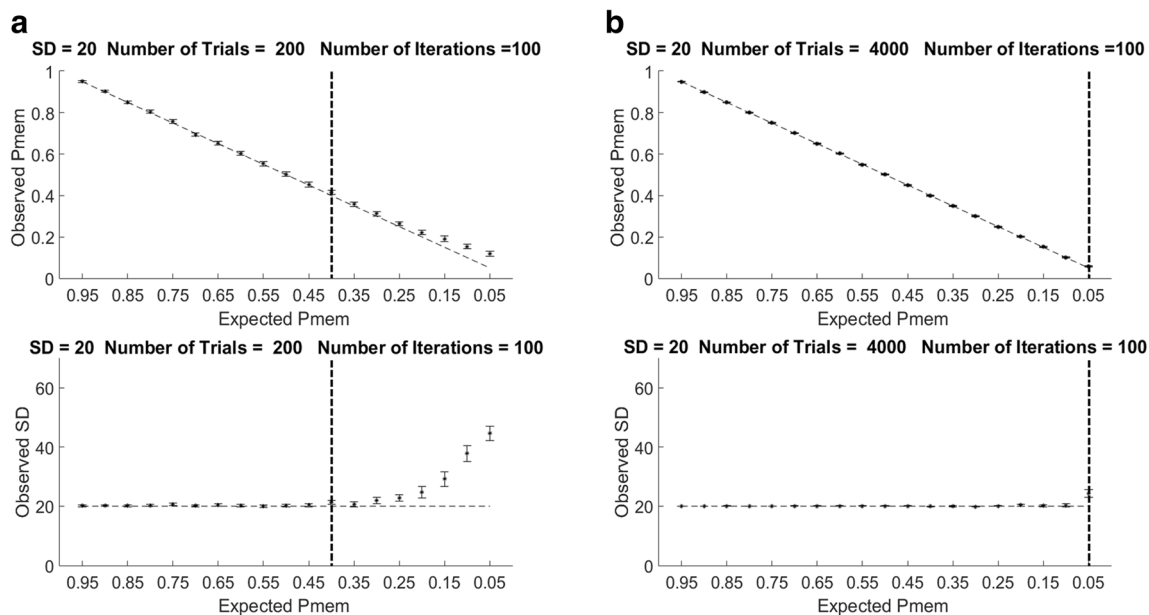


Fig. 3 Simulations of parameter estimates. **(A)** Parameter estimates obtained by taking 100 iterations of 200 trials each from distributions of known probability of retrieval (P_{mem}) and precision (SD). Error bars represent 95% confidence intervals, and dotted lines represent expected values. Top graph is observed probability of retrieval bottom is observed

precision. **(B)** Parameter estimates obtained by taking 100 iterations of 4,000 trials each from distributions of known probability of retrieval (P_{mem}) and precision (SD). Error bars represent 95% confidence interval, and dotted lines represent expected values. Top graph is observed probability of retrieval bottom is observed precision

(with the correct hue) was presented simultaneously with a color wheel and subjects were instructed to select the color of the presented item by clicking on the color wheel. The learning period was followed by a recall test of all of the items (with item-by-item feedback on the degree of response error).

Analyses

Similar to Experiment 1, we relied on an aggregate fit to assess the mnemonic precision for all subjects and then looked at individual fits for subjects who retrieved at least 40% of the items ($P_{\text{mem}} > 40\%$). Additional simulations with fewer trials revealed that this also was an appropriate cutoff for subjects who did not complete all trials (Figure S1).

Results

Aggregate

Seventy-four percent (CrI: -2.0% , $+1.5\%$) of the items were recalled during the initial test, and as expected, participants correctly selected responses for more than 99% (CrI: -0.4% , $+0.3\%$) of the items during the restudy task when the stimuli were physically present to guide responses. Not surprisingly, precision was substantially higher for the restudy ($SD = 7.2^\circ$, CrI: -0.2° , $+0.2^\circ$) than for the memory task (18.6° , CrI: -0.8° , $+0.6^\circ$).

At delayed test (Fig. 5), subjects recalled a significantly higher proportion of items that were initially

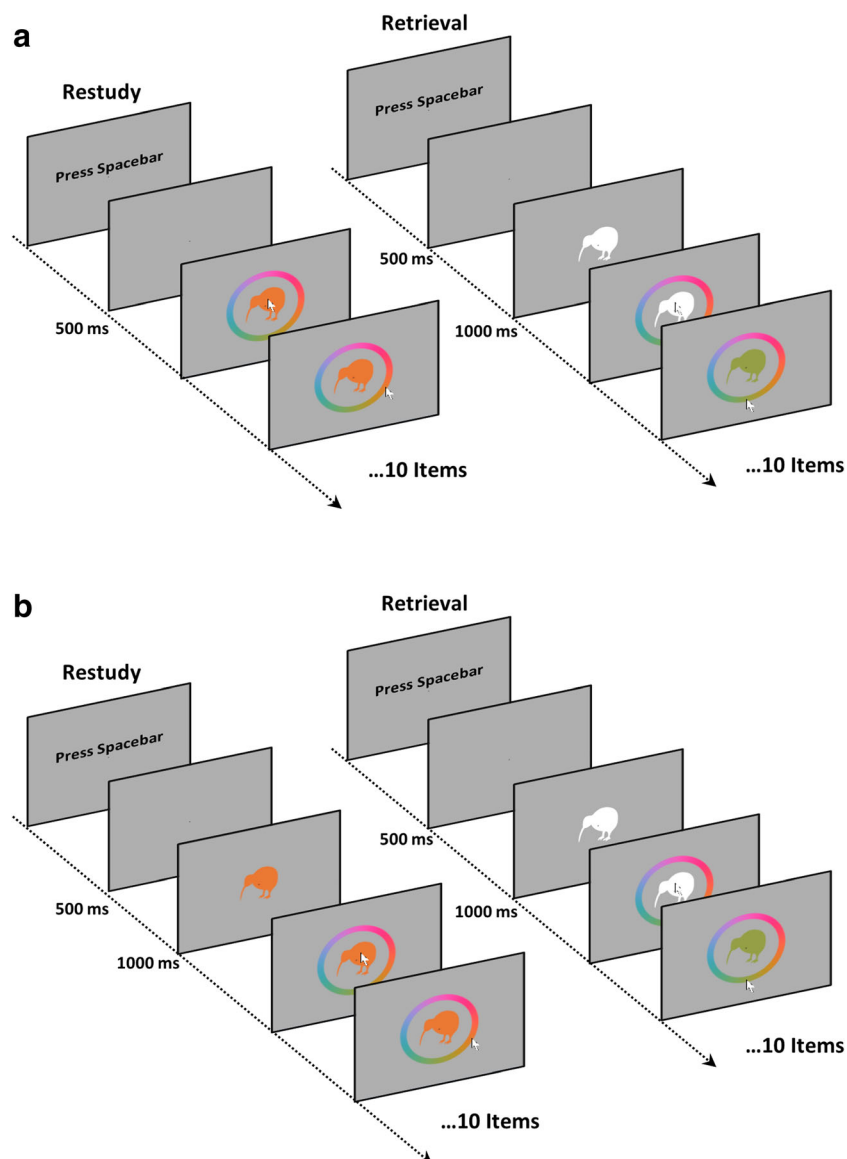


Fig. 4 Task diagram for Experiment 2+3. (A) Stimuli timing for restudy and retrieval for Experiment 2. (B) Stimuli timing for restudy and retrieval for Experiment 3. Final test timing was the same as initial test timing

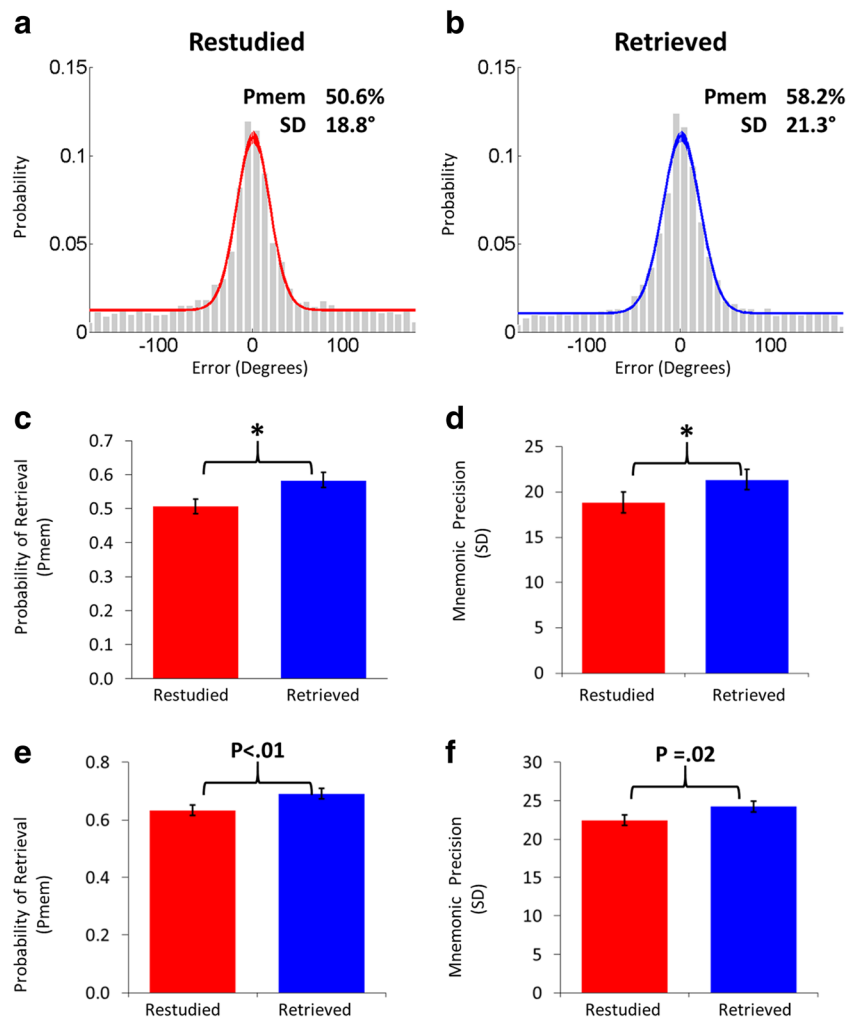


Fig. 5 Delayed test results from Experiment 2. **(A)** Aggregate fit of delayed response errors from all subjects for initially restudied items. **(B)** Aggregate fit of delayed response errors from all subjects for initially retrieved items. **(C+D)** Aggregate parameter estimates of probability of retrieval (P_{mem}) and mnemonic precision (SD) at delayed test*. Error bars represent Bayesian credibility intervals of the fits. **(E+F)**

Average of individual parameter estimates of P_{mem} and SD for subjects who successfully retrieved >40% of the items at the delayed test. Error bars represent 95% within-subjects confidence intervals that were calculated by normalizing to remove within subject variance (Loftus and Masson, 1994). *Denotes nonoverlapping credibility intervals

retrieved (58.2%, CrI: -2.0% , $+2.5\%$) than items that were restudied (50.6%, CrI: -2.1% , $+2.1\%$). Unlike the results of Experiment 1, mnemonic precision was significantly better for restudied items (18.8° , CrI: -1.2° , $+1.2^\circ$) relative to retrieved items (21.3° , CrI: -1.0° , $+1.2^\circ$; Fig. 5). As we report below, however, this relative disadvantage in mnemonic precision in the testing condition did not replicate in Experiment 3.

Individual parameter comparisons (Delayed Test)

Analysis of the subset of subjects who remembered at least 40% of items ($n = 17$) in both conditions revealed that subjects recalled a significantly higher proportion of items they had previously retrieved ($M = 69.1\%$, $SD = 14.8\%$) relative to

items that were previously restudied ($M = 63.4\%$, $SD = 14.4\%$, $t(16) = -3.06$, $p = 0.008$). As in the aggregate data, subjects exhibited superior mnemonic precision for items that they had previously restudied ($M = 22.5^\circ$, $SD = 8.2^\circ$) relative to retrieved items ($M = 24.3^\circ$, $SD = 9.2^\circ$, $t(16) = -2.60$, $p = 0.02$).

Discussion

As in Experiment 1, retrieval practice improved the probability of successful delayed recall but not mnemonic precision. Thus, the benefits of retrieval practice on probability of retrieval were robust when the control condition allowed extra time to restudy the memoranda. In Experiment 3, we tested whether a similar empirical

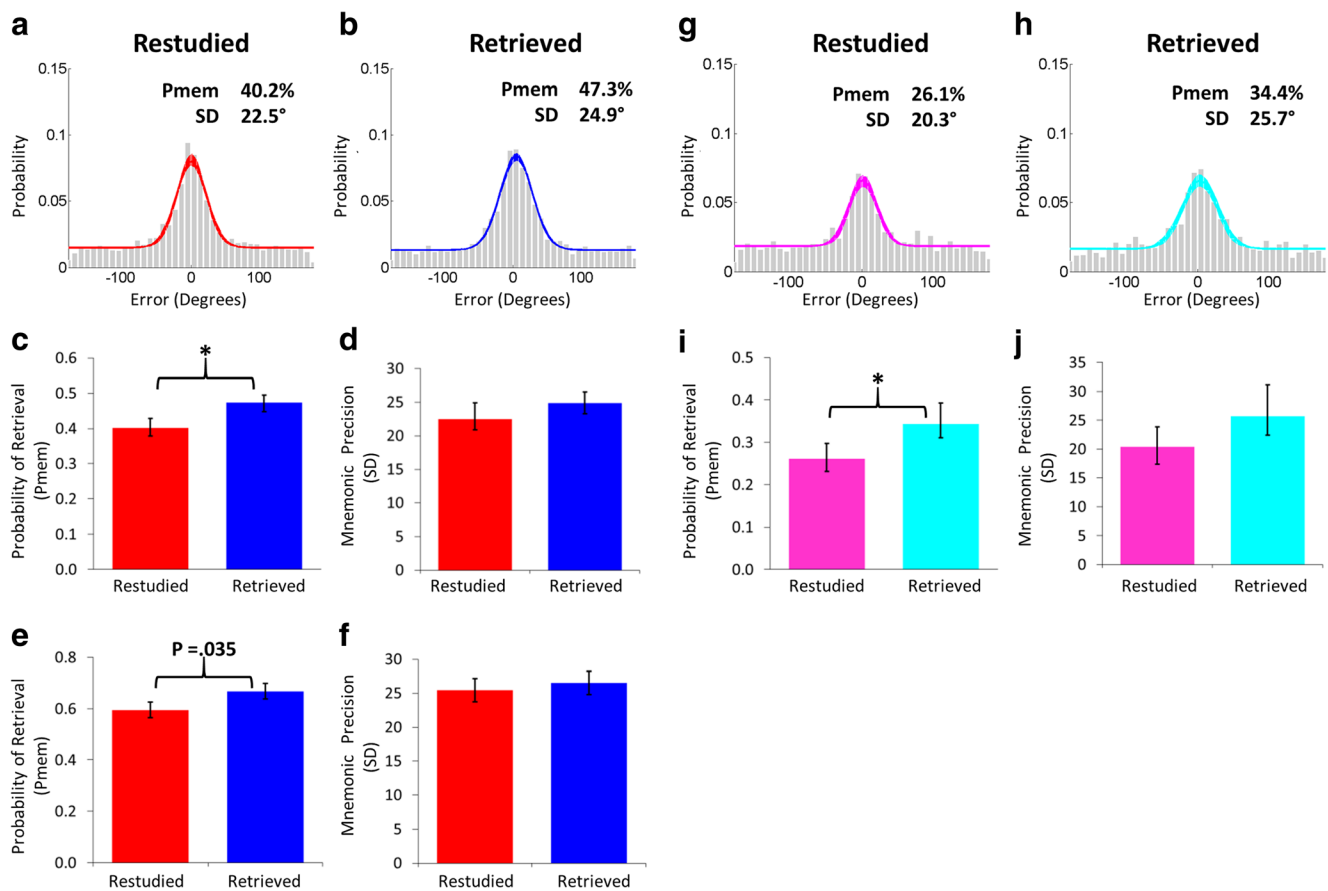


Fig. 6 Delayed test results from Experiment 3. **(A)** Aggregate fit of 30-min delayed response errors from all subjects for initially restudied items. **(B)** Aggregate fit of 30-min delayed response errors from all subjects for initially retrieved items. **(C+D)** Aggregate parameter estimates of probability of retrieval (P_{mem}) and mnemonic precision (SD) at 30-min delayed response errors. Error bars represent Bayesian credibility intervals of the fits*. **(E+F)** Average of individual parameter estimates of P_{mem} and SD for subjects who successfully retrieved >40% of the items at the 30-min delayed test. Error bars represent 95% within-subjects

confidence intervals that were calculated by normalizing to remove within subject variance (Loftus and Masson, 1994). **(G)** Aggregate fit of >24-hr delayed response errors from all subjects for delayed test of initially restudied items. **(H)** Aggregate fit of >24-hr delayed response errors from all subjects for delayed test of initially retrieved items. **(I+J)** Aggregate parameter estimate for P_{mem} and SD during >24-hr delay*. Error bars represent Bayesian credibility intervals of the fits. *Denotes nonoverlapping credibility intervals

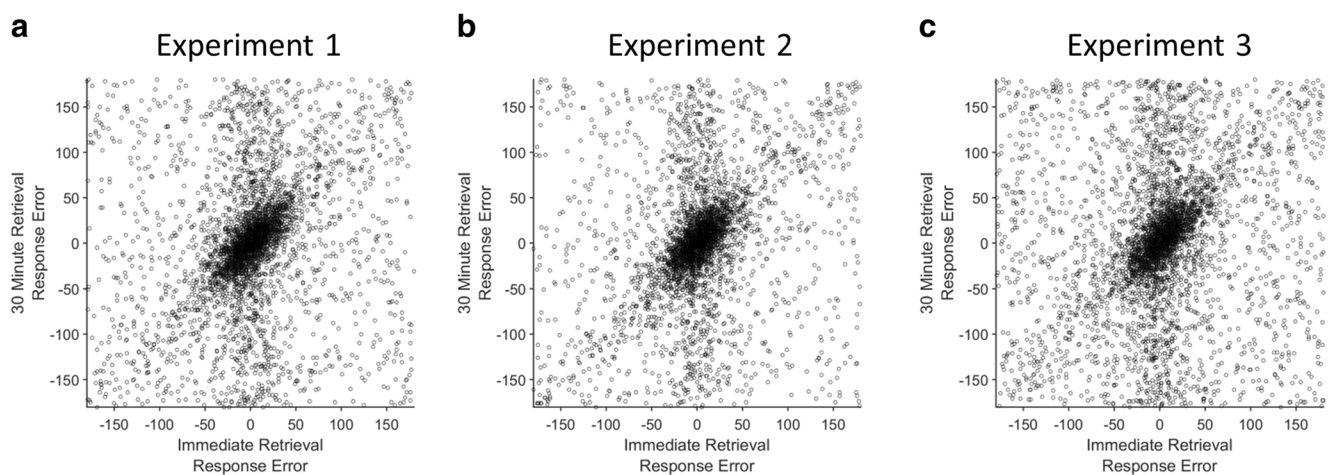


Fig. 7 **(A-C)** Scatter plot of immediate and 30-min delayed responses for retrieved items for all subjects in Experiments 1, 2, and 3, respectively. Responses from all subjects were pooled for the visualization

pattern would emerge when we equated the amount of exposure time between retrieval and restudy and whether the same pattern would emerge following a >24-hour retention interval.

Experiment 3

Method

Participants

Twenty-three students from the University of Oregon participated in Experiment 3 for course credit. Two participants who did not complete all trials in the time allotted were included in the experiment. All participants gave informed consent according to procedures approved by the University of Oregon institutional review board.

Task

The task in Experiment 3 was the same as Experiment 2 except for two differences. First, to equate total presentation time with that in the testing condition, restudied items were displayed for 1 s before subjects could respond (Fig. 4B). Second, to determine if the same pattern of results would emerge over a longer delay, the two runs of the task were completed on separate days, 1–4 days apart. This allowed for a >24-hr delayed retrieval of the items learned from the first run before subjects completed the second run of the experiment on day 2. Twenty subjects completed the surprise second retrieval period (3 subjects arrived late for the session and skipped the >24-hr retrieval to ensure a prompt finish).

Analyses

Analyses were identical to Experiments 1 and 2. Only the aggregate analysis was applied to the >24-hr delayed test, because subjects were only tested on 100 items in each condition and probability of retrieval was low.

Results

Aggregate

Sixty-three percent of the items (CrI: -1.8% , $+2.3\%$) were recalled during the initial test, and as expected, participants correctly selected responses for 99.7% (CrI: -0.2% , $+0.1\%$) of the items during the restudy task. Also as expected, precision was significantly higher for restudy ($SD = 7.7^\circ$, CrI: -0.2° , $+0.2^\circ$) than for retrieval (21.3° , CrI: -0.9° , $+1.1^\circ$).

At delayed test (Fig. 6), subjects recalled a significantly higher proportion of items that were previously retrieved

(47.3% , CrI: -2.6% , $+2.1\%$) than items that were previously restudied (40.2% , CrI: -2.2% , $+2.8\%$; Fig. 6). Unlike the results from Experiment 2, estimates of mnemonic precision were not significantly different for retrieved (24.9° , CrI: -1.6° , $+1.7^\circ$) and restudied (22.5° , CrI: -1.6° , $+2.5^\circ$) items. Thus, although Experiment 2 revealed relatively better precision in the restudy condition, this does not appear to be a robust empirical pattern.

The pattern of results observed during the test after more than 24 hr was similar to the pattern of results for the first delayed test. Subjects recalled a significantly higher proportion of items that they had previously retrieved (34.4% , CrI: -3.2% , $+5.0\%$) than items that they had restudied (26.1% , CrI: -3.1% , $+3.5\%$; Fig. 6). Estimates of mnemonic precision were not significantly different for retrieved (25.7° , CrI: -3.2° , $+5.5^\circ$) and restudied (20.4° , CrI: -3.0° , $+3.5^\circ$) items.

Individual parameter comparisons (Delayed Test)

Analysis of the subset of subjects who successfully retrieved at least 40% items ($n = 12$) revealed that subjects recalled a significantly higher proportion of previously retrieved items ($M = 66.8\%$, $SD = 11.4\%$) relative to previously restudied items ($M = 59.5\%$, $SD = 10.4\%$, $t(11) = -2.35$; $p = 0.039$). In contrast to the findings from Experiment 2, subjects exhibited similar precision for items that they had previously restudied ($M = 25.5^\circ$, $SD = 7.3^\circ$) relative to previously retrieved items ($M = 26.6^\circ$, $SD = 5.0^\circ$, $t(11) = 0.62$; $p = 0.54$).

Discussion

The findings from Experiment 3 are in line with the findings from Experiment 1 and Experiment 2. Retrieval practice improves the probability of successful delayed recall but does not improve mnemonic precision. Thus, the benefits of retrieval practice on recall probability were robust when the control condition allowed extra time to restudy the memoranda and when delayed recall did not take place for more than 24 hours.

General discussion

In three experiments, we demonstrated that retrieval practice improves probability of retrieval but not mnemonic precision. Furthermore, in Experiments 2 and 3 subjects provided a response to restudied items by selecting the color they were viewing on the color wheel. Thus, we were able to replicate a critical finding of Carpenter and Kelly (2012) that testing effects are still observed when subjects are required to make a response to restudied items. This line of results supports the idea that the benefits of retrieval practice are due to the act of retrieving information from long-term memory and not simply to subjects making a response for tested material but not for restudied material.

Ruling out a verbal code

One alternative explanation for the absence of an improvement in memory quality is that subjects relied on coarse verbal labels for color rather than a fine grained visual memory. For example, subjects could have remembered the name of a color and randomly responded within a section of the color wheel they associated with that label. Fortunately, the use of a continuous report measure allowed us to observe the relationship between an early retrieval attempt and a later retrieval attempt of the same item. This allowed us to calculate the correlation between response errors during initial test and response errors at delayed test for each subject. Indeed, the direction of error in the observers' responses in the immediate recall task predicted the direction of error in the delayed recall task. This bias was reliably different from zero for retrieved items in all experiments [Experiment 1 ($Z = 3.8, p < 0.001$), Experiment 2 ($Z = 4.1, p < 0.001$), and Experiment 3 ($Z = 4.2, p < 0.001$), see Fig. 7 for a visualization of aggregate responses]. These observations provide evidence against the use of a purely verbal code. If subjects were relying solely on a coarse verbal label and then guessing randomly within a section of the color wheel associated with that label when responding, we would expect early and late retrieval errors to be uncorrelated. However, we observed that delayed responses were biased by immediate responses. This bias suggests that subjects are able to maintain more than a coarse verbal representation of color. Furthermore, this is additional evidence for the idea that retrieval from memory is not a passive process and that subjects are reminded of and re-encode past representations of an item when given a test (Hintzman, 2011). This response bias finding dovetails with the episodic context theory of retrieval practice (Lehman, Smith, and Karpicke, 2014), which maintains that the context associated with a previously studied item is updated during the subsequent retrieval of that item to include features of both contexts. In this case, fine-grained prior reports can be added as another layer of context that is encoded when an item is later retrieved.

Conclusions

Extant models of the retrieval practice effect have asserted that testing enhances the accessibility of learned associations rather than the fidelity of the retrieved memories (Carpenter, 2009; Lehman, Smith, and Karpicke, 2014). The evidence for this assertion has been inconclusive, however, because of a heavy reliance on discrete word or picture stimuli that preclude a clear measure of item specific mnemonic precision. We measured performance in a test that required recall of colors from a continuous 360-degree space, and we used an analytic approach that enables distinct estimates of the probability of successful retrieval and the precision of the retrieved

representations. The results were clear at both the aggregate and individual subject levels. Retrieval practice selectively enhances the probability of recall without improving mnemonic precision. Thus, even though both accessibility and fidelity can determine memory performance, the selective effect of retrieval practice on the former highlights the utility of distinguishing these aspects of memory function.

Acknowledgments This work was funded in full by National Institutes of Health ROI-MH087214 to Edward Awh. The authors thank Anubhav Gupta and Dylan Sietz for help with data collection. Both authors conceived and designed the experiments and contributed to writing the manuscript. D.W.S. collected and analyzed the data.

References

- Brady, T. F., Konkle, T. F., Gill, J., Oliva, A., & Alvarez, G. A. (2013). Visual long-term memory has the same limit on fidelity as visual working memory. *Psychological Science*, *24*(6), 981–990.
- Brainard, D. H. (1997). The Psychophysics Toolbox. *Spatial Vision*, *10*, 433–436.
- Carpenter, S. K. (2009). Cue strength as a moderator of the testing effect: the benefits of elaborative retrieval. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *35*(6), 1563–1569.
- Carpenter, S. K., & Pashler, H. (2007). Testing beyond words: using tests to enhance visuospatial map learning. *Psychonomic Bulletin & Review*, *14*(3), 474–8.
- Carpenter, S. K., & Kelly, J. W. (2012). Tests enhance retention and transfer of spatial learning. *Psychonomic Bulletin & Review*, *19*(3), 443–8.
- Carrier, M., & Pashler, H. (1992). The influence of retrieval on retention. *Memory and Cognition*, *20*, 632–642.
- Chan, J. C. K., & McDermott, K. B. (2007). The testing effect in recognition memory: a dual process account. *Journal of Experimental Psychology Learning, Memory, and Cognition*, *33*(2), 431–7.
- Hintzman, D. L. (2011). Research strategy in the study of memory: fads, fallacies, and the search for the “coordinates of truth.” *Perspectives on Psychological Science*, *6*(3), 253–271.
- Lehman, M., Smith, M. A., & Karpicke, J. D. (2014). Toward an episodic context account of retrieval-based learning: Dissociating retrieval practice and elaboration. *Journal of Experimental Psychology Learning, Memory, and Cognition*, *40*(4), 1–8.
- Loftus, G. R., & Masson, M. E. J. (1994). Using confidence intervals in within-subject designs. *Psychonomic Bulletin & Review*, *1*(4), 476–490.
- Luck, S. J., & Vogel, E. K. (2013). Visual working memory capacity: from psychophysics and neurobiology to individual differences. *Trends in Cognitive Sciences*, *17*(8), 391–400.
- Pelli, D. G. (1997). The VideoToolbox software for visual psychophysics: Transforming numbers into movies. *Spatial Vision*, *10*, 437–442.
- Pyc, M. A., & Rawson, K. A. (2009). Testing the retrieval effort hypothesis: Does greater difficulty correctly recalling information lead to higher levels of memory? *Journal of Memory and Language*, *60*(4), 437–447.
- Roediger, H. L., & Karpicke, J. D. (2006). The Power of Testing Memory. *Perspectives on Psychological Science*, *1*(3), 181–210.
- Rohrer, D., Taylor, K., & Sholar, B. (2010). Tests enhance the transfer of learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *36*(1), 233.
- Suchow, J. W., Brady, T. F., Fougny, D., & Alvarez, G. A. (2013). Modeling visual working memory with the MemToolbox. *Journal of Vision*, *13*(1), 9.

- Szpunar, K. K., McDermott, K. B., & Roediger, H. L. (2008). Testing during study insulates against the buildup of proactive interference. *Journal of Experimental Psychology Learning, Memory, and Cognition*, *34*(6), 1392–9.
- Wheeler, M. A., & Roediger, H. L. (1992). Disparate Effects of Repeated Testing: Reconciling Ballard's (1913) and Bartlett's (1932) Results. *Psychological Science*, *3*(4), 240–245.
- Zhang, W., & Luck, S. J. (2008). Discrete fixed-resolution representations in visual working memory. *Nature*, *453*(7192), 233–5.