

Journal of Experimental Psychology: Learning, Memory, and Cognition

No Evidence of a Visual Testing Effect for Novel, Meaningless Objects

Anna C. McCarter, David E. Huber, and Rosemary A. Cowell

Online First Publication, February 13, 2025. <https://dx.doi.org/10.1037/xlm0001430>

CITATION

McCarter, A. C., Huber, D. E., & Cowell, R. A. (2025). No evidence of a visual testing effect for novel, meaningless objects. *Journal of Experimental Psychology: Learning, Memory, and Cognition*. Advance online publication. <https://dx.doi.org/10.1037/xlm0001430>

No Evidence of a Visual Testing Effect for Novel, Meaningless Objects

Anna C. McCarter¹, David E. Huber², and Rosemary A. Cowell^{2, 3}

¹ Department of Psychological and Brain Sciences, University of Massachusetts, Amherst

² Department of Psychology and Neuroscience, University of Colorado Boulder

³ Institute of Cognitive Science, University of Colorado Boulder



The testing effect is a well-established phenomenon in which memory is better for information that has been enhanced through practice tests rather than through restudying. However, this phenomenon has been studied almost exclusively with verbal or semantically meaningful material. We explored whether the testing effect holds for abstract visual material that lacks both meaning and verbal labels. In a series of six experiments, no evidence for a testing effect was found. Each experiment changed the nature of test practice in different ways that were designed to bolster test practice relative to restudy, such as imposing a delay before the final test, providing different kinds of choice options, providing different kinds of practice feedback, and using drawing as the form of test practice, and yet, the performance after test practice was either similar to the performance after restudy or in some cases significantly worse than restudy (i.e., a negative testing effect). We discuss the theoretical implications of these results, which suggest either that the testing effect relies on properties that our stimuli did not possess—for example, semantic content, high-dimensional content, or preexisting neocortical representations—or that eliciting a testing effect for visual material requires radically different task parameters than for verbal material.


Keywords: testing effect, visual memory, long-term memory

Supplemental materials: <https://doi.org/10.1037/xlm0001430.supp>

The testing effect is a well-established phenomenon in which long-term retention is better for information that has been learned through practice tests rather than through restudy (Abbott, 1909; Roediger & Karpicke, 2006a, 2006b). Despite the robust nature of testing effects and the substantial testing effect literature, there is considerable debate over the underlying mechanisms (Rowland, 2014). For instance, some theories assume that retrieval practice benefits arise from changes in the semantic content of memories, whereas other theories suppose that retrieval practice benefits arise

from more general processes. If the testing effect reflects more general processes, a benefit for testing as compared to restudy should occur even for material that is devoid of meaning (e.g., abstract visual material). However, the testing effect has been examined almost exclusively with verbal, meaningful stimuli. Thus, it remains unclear whether test practice for meaningless visual stimuli can produce stronger, longer lasting memories than restudy of meaningless visual stimuli. To address this question, we report six experiments using visual recall practice with novel, meaningless visual objects. To

Neil W. Mulligan served as action editor.

Anna C. McCarter  <https://orcid.org/0000-0002-1107-2494>


The data sets and additional relevant materials generated during the present study are available at <https://osf.io/57jsv/> (McCarter, 2023). This work was presented at the University of Massachusetts at Amherst 2022 Interdisciplinary Neurosciences Conference, the 2022 Context and Episodic Memory Symposium, the 2022 Object Perception and Memory Symposium, the 2022 Psychonomic Society Annual Meeting, and the 2023 Vision Sciences Society Annual Meeting. An abstract describing this work was published as a conference proceeding in the *Journal of Vision*. A preprint of this work has been posted on the Open Science Framework.

This work was supported by the National Institute of Health (Grant IRF1MH114277-01 to Rosemary A. Cowell and David E. Huber) and a University of Massachusetts at Amherst Predissertation Grant to Anna C. McCarter. The authors thank Aayush Patel, Aisling Finnegan,

Colleen Dunn, Hannah Laird, and Max Kozlowski for their help with data collection.

Anna C. McCarter contributed to conceptualization, data curation, formal analysis, funding acquisition, investigation, methodology, project administration, resources, software, validation, visualization, writing—original draft, and writing—review and editing. David E. Huber contributed to conceptualization, funding acquisition, methodology, resources, supervision, and writing—review and editing. Rosemary A. Cowell contributed to conceptualization, funding acquisition, methodology, resources, supervision, and writing—review and editing.

 The data are available at <https://osf.io/57jsv/>.

 The experimental materials are available at <https://osf.io/57jsv/>.

Correspondence concerning this article should be addressed to Anna C. McCarter, Department of Psychological and Brain Sciences, University of Massachusetts, Amherst, 135 Hicks Way, Amherst, MA 01003, United States. Email: acmccarter@umass.edu

preview our results, in no case did we find evidence of a testing effect for these abstract visual stimuli.

Three closely related theories assume that the testing effect reflects the integration and strengthening of representations within semantic knowledge. First, the elaborative retrieval hypothesis suggests that testing is more beneficial than restudy because testing leads to the creation and strengthening of related information, which ultimately provides a greater number of pathways (i.e., via alternative retrieval cues) to the target (Carpenter, 2009). Second, the mediator effectiveness hypothesis claims that the testing effect arises from the strengthening of semantic connections between the cue and the retrieved target (Pyc & Rawson, 2010). Third, fuzzy trace theory suggests that restudy emphasizes surface parts whereas testing emphasizes the broader, semantic information, with the latter being more beneficial for long-term memory (Bouwmeester & Verkoeijen, 2011; Reyna & Brainerd, 1995). All three of these theories suggest that semantic content is essential for the testing effect. In addition, Antony et al. (2017) put forward a theory that is not explicitly couched in terms of semantic content but shares elements with the three semantic theories. It proposes that retrieval acts similarly to offline consolidation during sleep, enhancing the integration of new memories into preexisting knowledge structures in the neocortex (i.e., semantic networks), helping differentiate overlapping memories, and shifting the dependence of new memories from the hippocampus to the neocortex. Like the semantic content-mediated theories, this account predicts a greater testing effect for information already represented in neocortical knowledge schemas than for unlearned material like abstract visual stimuli or nonsense syllables.

In contrast to semantic accounts of the testing effect, several other theories suppose that the benefits of retrieval practice should occur regardless of the modality or nature of the practiced material. For instance, transfer-appropriate processing (C. D. Morris et al., 1977) simply states that a practice test should be the best way to learn material to the extent that the practice test is well-matched to the subsequent final test in terms of the processes involved. The primary and convergent retrieval model of recall (Hopper & Huber, 2018) can be considered a specific instantiation of transfer-appropriate processing. This model assumes that the pattern completion process of recall affords a unique opportunity to create associations within the to-be-remembered stimulus (e.g., from the first letter of a name to the remaining letters), which enhances subsequent recall performance. Providing a general account of retrieval practice, the desirable difficulties framework (Bjork & Bjork, 2011) assumes that the effort necessary for a practice test enhances learning. In a similar vein, there is evidence that prediction errors that occur on practice tests are beneficial to memory (Mozer et al., 2004). Finally, the episodic context account (Karpicke et al., 2014) assumes that the benefits of retrieval practice arise from the retrieval and updating of the learning context associated with the remembered stimuli. Thus, according to these theories, the nature of the to-be-remembered material should not matter because the benefits arise from the act of recall (transfer-appropriate processing), the effort and errors that come with recall (desirable difficulties and prediction errors), or the association of the recalled item with the practice test context (episodic account).

There have been several studies examining the testing effect with visual content, although nearly all prior studies used stimuli that contained meaning or verbal labels of some kind. For instance, Kang (2010) found a testing effect for learning the association between

English words and Chinese characters. Carpenter and Pashler (2007) found a testing effect for meaningful landmarks on a map (e.g., practice retrieving the missing item on the map). Sutterer and Awh (2016) and Schuetze et al. (2019) found a testing effect for learning novel associations between the outline of common objects and their fill color. Providing additional examples, a testing effect for visual stimuli has been found for face–name pairs (Carpenter & DeLosh, 2005; P. E. Morris & Fritz, 2000; Tse et al., 2010), the spatial layout of objects (Carpenter & Kelly, 2012), Adrinka symbol–word pairs (Coppens et al., 2011), bird image and family name pairs (Siler & Benjamin, 2020), scene–object pairs (Jonker et al., 2018), and word–image pairs (Ferreira & Wimber, 2023; Lifanov et al., 2021). Importantly, all of these experiments involved meaningful, semantic stimuli in some manner either by presenting known visual objects (e.g., faces, birds, map locations) or by involving the association of meaningless visual objects (e.g., Chinese characters, Adrinka symbols) with known words. So, it is unclear if these examples reflect retrieval practice benefits for purely visual information or whether these testing effects were mediated by semantic learning.

To address the question of whether the visual testing effect occurs only for meaningful visual objects, Ferreira and Wimber (2023) ran four experiments in which the visual object stimuli were meaningless squiggles for two experiments but well-known objects in two other experiments using a similar procedure. In every experiment, participants initially learned to associate an unrelated word with each of the visual objects. Then, in a between-subjects manipulation, additional practice occurred either by restudying the word–object pairs for 7.5 s or by attempting to vividly imagine the object in response to the cue word for 5 s, followed by 2.5 s of viewing the correct object, with this final 2.5 s of viewing providing visual feedback for the imagination task. Thus, both restudy and test practice involved viewing the objects (either for 7.5 s or 2.5 s) but only in the test practice condition did participants attempt to imagine the object. The final test was a three-alternative forced choice between objects in two experiments and remember/know/new responses to single objects in the other two experiments. Supporting the hypothesis that visual testing effects only occur for meaningful objects, the two experiments that used meaningless squiggles failed to find any advantage for test practice over restudy and even found a negative testing effect in one condition. In contrast, one condition in one of the two experiments with meaningful objects found a positive testing effect.

Although the study by Ferreira and Wimber (2023) provides some indication that testing effects fail to occur with meaningless visual objects, there are some limitations to the study. First, the test practice condition in the Ferreira and Wimber study *did not require any behavioral response*. Rather than spending 5 s imagining the associated object, some participants may have decided to wait 5 s (i.e., without attempting to visually recall) and then use the 2.5 s of feedback to engage in restudy. If some participants adopted such a strategy, this would reduce any differences between the restudy and test practice conditions. In our experiments, we gave participants a forced choice test after the visual retrieval period and, critically, *we did not present the retrieval cue during this forced choice test* in any of the experiments except Experiment 3. If participants did not pay attention and/or attempt visual retrieval during the visual imagination period for Experiments 1, 2, and 4, they would be at a complete loss on this test (Experiment 5 did not present the retrieval cue for the forced choice, but the task could be accomplished

without paying attention to the retrieval cue for other reasons, explained below). Thus, the retrieval practice task in Experiments 1, 2, and 4 should motivate participants to engage in visual retrieval, and, more importantly, it provides a measure of their effort (i.e., if they did not pay attention during the visual imagination period, then the performance on the practice test would necessarily be at chance). In Experiment 6, we took things a step further by requiring that participants draw the contents of their visual retrieval during retrieval practice.

A second key limitation of the Ferreira and Wimber study is that words were used as one half of each stimulus pair (with meaningless visual stimuli comprising the other half). Words are inherently meaningful, which undermines any claim that this paradigm examines the testing effect for meaningless content. The inclusion of words for all practice pairs may have encouraged participants to extract some meaning from the novel meaningless shapes to associate with the words (e.g., “that outline sort of looks like an antelope and to help me remember that the shape goes with the word moon, I’ll imagine an antelope on the moon”), although such a semantic strategy will obviously be more effective when the to-be-paired stimuli are meaningful (e.g., if there were a picture of an antelope, then it would be easier to learn this semantic association). In contrast, our experiments presented meaningless shapes in isolation and retrieval practice concerned memory for visual aspects of a single meaningless object, rather than how the object might or might not relate to a word. By forcing participants to make visual associations (e.g., the contour of the shape and the nature of the fill pattern), we can ask whether visual retrieval practice is an effective form of practice even for meaningless stimuli.

The present study does not attempt to determine whether the visual testing effect only occurs for meaningful visual objects—this can be determined only with statistical tests that directly compare meaningful and meaningless objects. Instead, the present study asks whether it is possible to obtain a visual testing effect for meaningless visual objects in a situation that is more likely to produce such an effect (i.e., a more heavy-handed manipulation of visual recall practice). Similar to the Ferreira and Wimber (2023) study, we asked participants to mentally imagine previously studied meaningless objects in response to a retrieval cue. However, unlike the Ferreira and Wimber study, the retrieval cue was a visual aspect of the object (either the outline shape or its fill pattern) rather than a word, and the task was to imagine the missing visual aspect (e.g., if shown the outline, attempt to imagine the fill pattern). The absence of any word stimuli should focus learning and retrieval on purely visual aspects of the stimuli, particularly given that the visual objects were meaningless. Crucially, after recall practice through visualization, participants were given a forced choice between two possible answers (e.g., two possible fill patterns), and *this forced choice was the very same forced choice that would later appear on the final test*. This practice of the same test as the final test should maximize the transfer-appropriate processing benefit that might arise from test practice (C. D. Morris et al., 1977). After each practice test trial, they were given explicit feedback about the correct choice, which should support accurate responding on the final test.

Thus, our study attempts to establish a testing effect for meaningless visual materials by including at least three important features, never previously combined: We use purely abstract, novel stimuli (rather than combinations of abstract items with meaningful

items); we employ an overt test of memory performance *during retrieval practice*; and, as opposed to the item-to-item associative learning typically used in testing effect studies, we use visual part-to-part learning that is more likely to engage low-level, visual representations.

Prior work has shown that whole-object study can promote part-to-part visual learning, that is, from one part of the object to another part (Sadil et al., 2019). The question asked in the present study was whether learning part-to-part visual associations via repeated retrieval of one visual part (e.g., fill pattern) in response to a different visual part (e.g., outline shape) would afford stronger learning than simply restudying the object. After the initial study of the objects, half of the objects received additional learning through a cued recall (part-to-part retrieval) practice task that included feedback, while the other half received additional learning through restudy. This was followed by a final cued recall test and two different recognition tests for all objects. In designing Experiment 1, we fully expected to find a testing effect and to avoid the criticism that the testing effect emerged from choosing an ineffective form of restudy, we used liking judgments as a form of restudy. Liking judgments are known to be a powerful form of incidental learning (Nairne et al., 2008), and, furthermore, liking judgments have been used in other testing effect studies (Congleton & Rajaram, 2012). After failing to find a testing effect in Experiment 1, we gradually tweaked the paradigm in ways that we believed would elicit a testing effect in a progressively more heavy-handed manner (an overview of all six experimental methods appears in Figure 1). In every case, the restudy condition was either similar to or significantly better than the visual cued recall practice condition, for both the final cued recall and the final recognition tests.

Experiment 1

Participants


































We assumed that the restudy condition produces average performance of 75% and the test practice condition produces average performance of 85%. This 10% testing effect advantage corresponds to a Cohen’s d of .5 for a within-subjects design with eight trials per condition, which is a “medium” effect size. A power analysis indicated that 25 participants would be needed to achieve 80% power to detect a testing effect (data were generated from binomial sampling, with eight trials per condition per participant, and assessed with a one-tailed dependent samples t test). The 25 participants (18 female, six male, one no response) ranged in age from 18 to 45 years ($M = 20.6$ years). All participants in all experiments were recruited from the University of Massachusetts at Amherst student population and received course extra credit for participating. This study was approved by the University of Massachusetts at Amherst Institutional Review Board (Protocol No. 1022), and participants each provided informed consent. All experiments were programmed and run using PsychoPy (Peirce et al., 2019).

Method

Materials

Sixteen different meaningless visual objects were created by combining one of eight different outline shapes with one of eight

Figure 1
Summary of Experimental Method for Each Experiment

	Key Manipulations	Practice Test Task	Practice Test Feedback	Re-Study Task	Additional Familiarization	Delay
Exp. 1	N/A	View Cue  2AFC Features  	Whole Object 	Likeability Rating  1 2 3 4 5	6 Rounds	Face Test
Exp. 2	2-7 day delay	View Cue  2AFC Features  	Whole Object 	Likeability Rating  1 2 3 4 5	6 Rounds	2-7 Day Delay & Face Test
Exp. 3	2AFC with whole objects	View Cue  2AFC Objects  	Whole Object 	Likeability Rating  1 2 3 4 5	2 Rounds	Face Test
Exp. 4	No whole object exposure in either condition	View Cue  2AFC Features  	Correct Feature  	View Cue & Correct Choice   	6 Rounds	Face Test
Exp. 5	New object-family assignment	View Cue  2AFC Features  	Correct Feature  	Likeability Rating  1 2 3 4 5	10 Rounds	Face Test
Exp. 6	Drawing in practice condition	View Cue & Draw Corresponding Feature  	Whole Object 	Likeability Rating  1 2 3 4 5	6 Rounds	Face Test

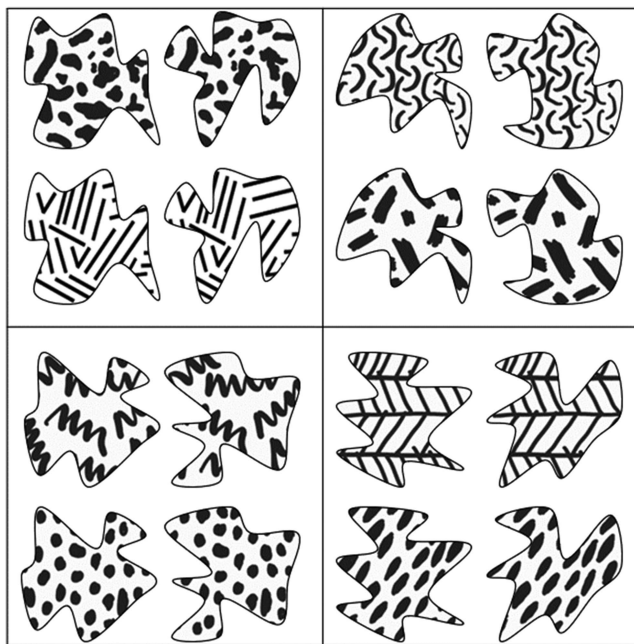
Note. All experiments used whole object study prior to additional familiarization (not shown). Additional familiarization comprised alternating blocks of practice tests and restudy, with half of the objects assigned to practice and the other half to restudy; each pair of two successive blocks—one study, one test—corresponded to one “round” of additional familiarization. Except for the order of block alternation and the assignment of stimuli to conditions, the design was fully within subjects. Except for Experiment 4, all experiments used incidental learning for restudy by having participants rate how much they liked the meaningless visual objects. The figure shows the presentation of an outline shape for cued recall of a fill pattern, but half of the blocks of familiarization presented a fill pattern for cued recall of an outline shape. Final cued recall tested all objects in both directions (from shape-to-fill and fill-to-shape). This was followed by two different recognition tests for all objects. After additional familiarization, all experiments presented the Cambridge Face Test (Duchaine & Nakayama, 2006) as a filler task before the final cued recall and recognition tests. The summary of “key manipulations” always refers to changes from Experiment 1. 2AFC = 2-alternative-forced-choice; Exp. = experiment; N/A = not applicable. See the online article for the color version of this figure.

different fill patterns (see Figure 2). Outline shapes were created by hand using the “curve” function in Microsoft PowerPoint, based on 12 randomly selected inflection points. As seen in Figure 2, this resulted in outline shapes with approximately five convex elements and five concave elements. The fill patterns were found through Google Images and selected for the property of being not easily named or labeled. As shown in Figure 2, the objects were divided into four families (one family is shown in each quadrant of Figure 2), with each family having two outline shapes (the columns of each family within the figure) and two fill patterns (the rows of each family within the figure) put together in different combinations to create four objects per family. The pair of objects along one diagonal in each family were used as to-be-learned targets, while the pair along the other diagonal served as lures. To counterbalance the stimuli across participants, two of the four families were randomly assigned to the restudy condition, and the other two were assigned to the test practice condition. This yielded four target objects assigned to the practice condition and four target objects assigned to the restudy condition.

Procedure

The stages of the experiment included an initial study of the eight objects, additional familiarization involving test practice and restudy, a filler face task, and three final tests (Figure 3). In Experiment 1, there were six rounds of additional familiarization, with each round composed of four trials of practice and four trials of restudy such that each of the target objects was familiarized once during each round.

The initial study stage presented all eight target objects in randomized order. Each target object was presented for 5 s, and then, participants were asked to rate how much they liked the object on a scale of 1–5 (Figure 3). Participants were informed prior to the initial study that they would later be tested on their memory for these objects. Additional familiarization involved four trials of test practice and four trials of restudy per round such that all eight target objects received one trial of additional familiarization per round. The order of test practice and restudy was randomized between participants. Restudy trials were identical to the initial study trials,

Figure 2*The 16 Novel, Meaningless Objects Used for All Experiments*

Note. In this figure and in all others depicting the stimuli, the exact stimulus outlines seen by participants are displayed, but for copyright reasons, only approximations of the patterns seen by participants are shown (see <https://osf.io/57jsv/> for full original materials). Families of four objects were created by combining two fill patterns (rows) with two outline shapes (columns). One family appears in each quadrant of the figure. Two objects from each family were the to-be-learned target objects and the other two were presented as lures during recognition (i.e., intact vs. rearranged recognition). For each participant, two of the four families were randomly assigned to the restudy condition, and the other two were assigned to the practice test condition.

presenting each target object for 5 s followed by a liking judgment. Liking judgments have been shown to elicit good levels of encoding (Hyde & Jenkins, 1969; Nairne et al., 2008; Whiffen & Karpicke, 2017). We chose to include this liking judgment in order to ensure that participants were looking at the screen and engaged. If we had simply told participants to restudy the item, they might have disengaged from the task.

Practice trials presented one part (either outline shape or fill pattern) as a retrieval cue of the target object. The retrieval cue was presented for 5 s with instructions to imagine the corresponding part (e.g., if shown the outline shape, imagine the corresponding fill pattern). After this covert visual recall period, participants were given an unspeeded 2-alternative-forced-choice (2AFC) between either the two shapes from that family or the two patterns from that family, depending on whether the cue was a pattern or shape, respectively. Participants were told to pick the part that they had imagined using a key press. Their choice was registered by placing a box around their choice; the box was green if correct and red if incorrect.

In addition to the appearance of the colored box, the correct whole target object was presented, and these three elements (e.g., the selected part with the colored box, the nonselected part, and the correct whole object) remained onscreen for 2 s (see Figure 3). In

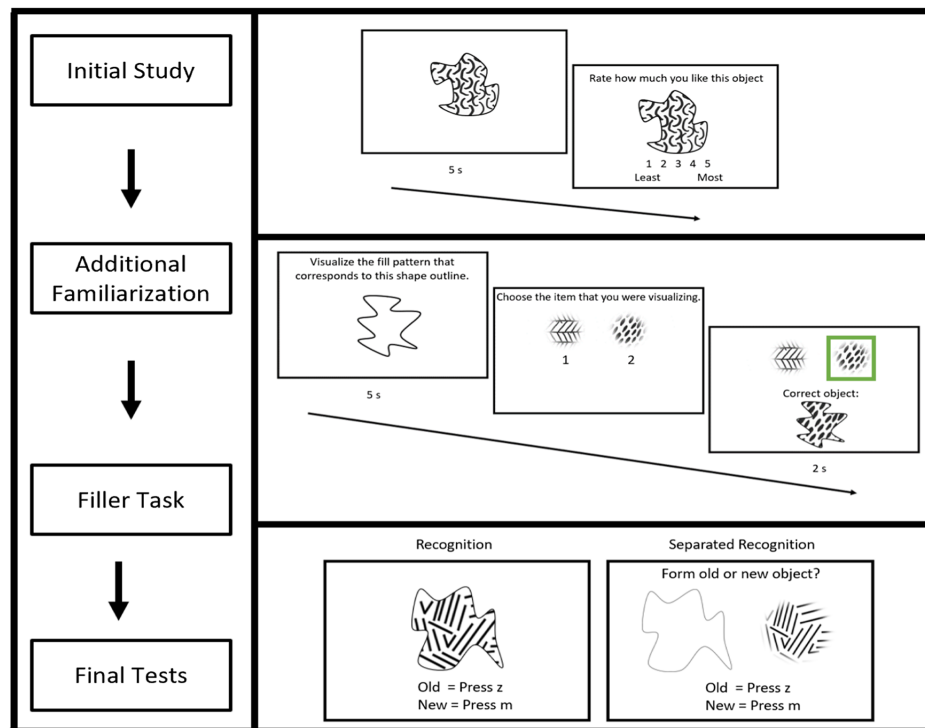
other words, not only were participants given a practice test but they were given a chance to restudy the object for 2 s as a form of feedback. In the absence of test practice feedback, restudy familiarization can produce higher performance than test practice familiarization for an immediate final test because test practice tends to benefit only those items that are correctly recalled (Jang et al., 2012; Kornell et al., 2011). However, this “bifurcation” in the performance between retrieved versus nonretrieved items should not apply in the current paradigm because we provide extensive correct-answer feedback in the retrieval condition. It was therefore expected that test practice would produce a relative advantage compared to restudy, even for an immediate final test. Because the benefits of cued recall practice can be directional, from cue to target (Aenugu & Huber, 2021), participants underwent both shape-to-fill and fill-to-shape practice for all objects, such that associations were learned in both directions. Within each test practice block, each object was tested in one direction (either shape-to-fill or fill-to-shape), but the direction for each object alternated across successive rounds of familiarization so that the two directions were seen an equal number of times. The tested direction of the first round was randomly determined for each object in each participant.

Following additional familiarization, participants were given the Cambridge Face Test (Duchaine & Nakayama, 2006) as a visually distracting filler task. This test took about 10–15 min to complete and was used to prevent maintenance of the objects in visual working memory. After the filler task, participants underwent three separate tests of their memory for the target objects. The first test was cued recall that was identical in format to the practice tests (covert retrieval followed by 2AFC) except that there was no feedback, and all eight target objects were tested (i.e., the restudy objects were also tested). Because both directions were tested (from shape-to-fill and fill-to-shape) separately for each object, on separate test trials, there were 16 trials of cued recall. The order of the 16 test trials was randomized for each participant.

Cued recall was followed by two different kinds of recognition that allowed separate examination of whether test practice preferentially strengthened integrated whole object representations (the first recognition test) versus shape–fill associations (the second recognition test). The first recognition test was whole object recognition. Participants saw whole objects one at a time and indicated whether each object was old or new (Figure 3). Each of the eight intact studied target objects was shown once and each of the eight rearranged lure objects was shown once, resulting in 16 recognition trials. No feedback was provided. This was followed by a second recognition test, termed “separated recognition.” This test was conceptually the same as the whole object recognition, with eight intact targets and eight rearranged lures, except that the parts were presented separately, side by side, instead of as a unified object (Figure 3).

One goal of the separated recognition test was an examination of recognition latencies, and so, each of the 16 shape–fill pairs was tested four times each (two with fill pattern on the left and outline shape on the right and two vice versa), providing 64 separated recognition trials. This goal was established in light of recent results indicating that even when test practice fails to produce higher accuracy than restudy, it can nevertheless produce faster responses (Hopper & Huber, 2018, 2019). For separated recognition, participants were instructed to respond as quickly as possible and a variable intertrial interval of 0.5–1.0 s was used to prevent them from preparing a response before viewing the test item.

Figure 3
Overview of Paradigm for Experiment 1



Note. Left: All eight target objects were viewed once during the initial study, followed by six rounds of additional familiarization, a filler task, and three final tests. Top right: Sample study/restudy trial from Experiment 1. Participants viewed an object for 5 s and then rated how much they like the object on a scale of 1–5. Middle right: Sample practice trial from Experiment 1. Participants viewed the cue and imagined the corresponding part for 5 s and then picked the part they were visualizing, followed by 2 s of feedback that included presentation of the correct whole object. The cue could be either the shape outline (as shown here) or the fill pattern. Bottom right: Example recognition and separated recognition trials. In the recognition test, participants saw the whole object and indicated with a keypress if it was old or new. In the separated recognition test, participants viewed a fill pattern and an outline shape and indicated with a keypress if it formed an old or new object. See the online article for the color version of this figure.

Transparency and Openness

We report how we determined our sample size, all data exclusions, all manipulations, and all measures in the study. All data, analysis code, and research materials are available at <https://osf.io/57jsv/>. Data were analyzed using R (R Core Team, 2021). The design and analysis of this study were not preregistered.

Results

The performance on the forced choice tests during test practice and on the final cued recall test was assessed with an analysis of average percent correct. The performance on yes/no recognition tests was assessed using equal variance signal detection theory (d') to control for response bias (Macmillan & Creelman, 2004). In calculating d' , the Hautus correction was used to avoid values of infinity that would otherwise emerge if either the hit or false alarm rate were zero or one (Hautus, 1995). In addition to accuracy, reaction times for the separated recognition trials were examined to assess any speed/accuracy trade-offs and in light of results indicating that test practice

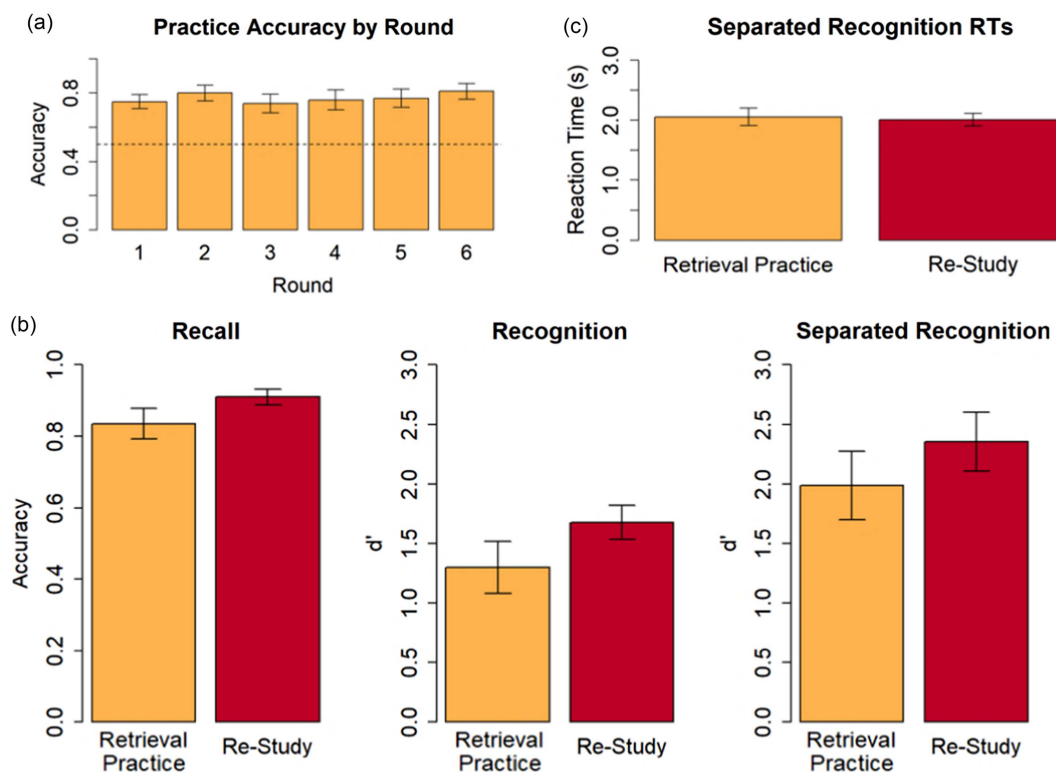
promotes faster responses (Hopper & Huber, 2018, 2019). The reaction time data were analyzed only for hits.

Participants performed well above chance on the practice task (Figure 4a), although somewhat surprisingly, there was no reliable evidence of improvement across blocks according to a linear trend analysis, $t(24) = 0.71$, $p = .242$. As seen in Figure 4b, there was no significant difference between practiced objects and restudied objects in recall, $t(24) = 1.90$, $p = .070$; recognition, $t(24) = 1.91$, $p = .069$; and separated recognition, $t(24) = 1.24$, $p = .227$. Numerically, there was a restudy benefit for all three of the final tests. In addition to accuracy and d' on the final memory tests, the reaction times for the separated recognition test were compared (Figure 4c). There was no significant difference between reaction times for practiced objects and restudied objects, $t(24) = 0.35$, $p = .727$. Plots of individual participant performance can be found in Supplemental Figure S1.

Discussion

Despite the robust prior evidence of testing effects in a variety of visual stimuli, we found no evidence of a purely visual testing

Figure 4
Results of Experiment 1



Note. (a) Practice accuracy by round of additional familiarization for Experiment 1. The dashed line represents chance-level performance. (b) Recall accuracy, recognition d' , and separated recognition d' for Experiment 1. There was no significant difference between practice and restudy performance in any of the tests. (c) Correct reaction times for intact shape/fill separated recognition pairs (i.e., reaction times for hits) in the separated recognition test for Experiment 1. There was no significant difference between practiced and restudied objects. All error bars represent standard error of the mean. RTs = reaction times. See the online article for the color version of this figure.

effect in Experiment 1. Participants performed worse on the practiced objects than the restudied objects in the final tests. Even the reaction times did not provide evidence of a practice benefit. In the experiments reported next, we made progressive changes to the paradigm in hopes of uncovering a purely visual testing effect.

Experiment 2

The goal of Experiment 2 was to see if adding a multiday delay between additional familiarization and the final tests would elicit a practice benefit. In the literature, it has been reported on several occasions that there is a stronger testing effect if the final tests are administered a few days after the learning (Coppens et al., 2011; Kang, 2010; Roediger & Karpicke, 2006a). However, it is important to note that, in paradigms that provide correct-answer feedback during retrieval practice (like Experiment 1), a testing effect is typically observed at all delay lengths, including immediate final test; this finding is intuitive because, with correct-answer feedback, participants have an opportunity for both retrieval and restudy in the retrieval practice condition giving retrieval practice a

clear advantage. Nonetheless, perhaps the short delay in Experiment 1—only 10–15 min between additional familiarization and the final tests—contributed to attenuating any testing effect that might otherwise have been observed.

Participants

Twenty-six people participated, five of whom were excluded for failing to return for the second day of experimentation and one of whom was excluded for accidentally closing the experiment before it was completed. Of the remaining 20 participants (19 female, one male), ages ranged from 18 to 23 years ($M = 19.8$ years).

Method

The methods for Experiment 2 were identical to the methods of Experiment 1 except for adding a multiday delay between additional familiarization and the final tests. Participants had a 2- to 7-day delay ($M = 4.0$ days) after they completed the additional familiarization and before the filler task and final tests.

Results

Participants gradually improved on the practice task across the six rounds of additional familiarization (Figure 5a) with a significant increasing trend per the linear analysis, $t(19) = 4.37, p < .001$. Considering that the test practice session of Experiment 2 was identical to Experiment 1, this suggests that the failure to find a reliable benefit in Experiment 1 was a Type II error. However, even though this demonstrates that test practice was beneficial, there was no evidence that test practice was more beneficial than restudy: As seen in Figure 5b, there was no significant difference between practiced objects and restudied objects in recall, $t(19) = 0, p = 1$; recognition, $t(19) = 0.08, p = .936$; and separated recognition, $t(19) = 0.69, p = .496$. The reaction times for the separated recognition test are shown in Figure 5c. There was no significant difference between the reaction times for practiced objects and restudied objects, $t(19) = 0.98, p = .338$. Plots of individual participant performance can be found in Supplemental Figure S2.

Discussion

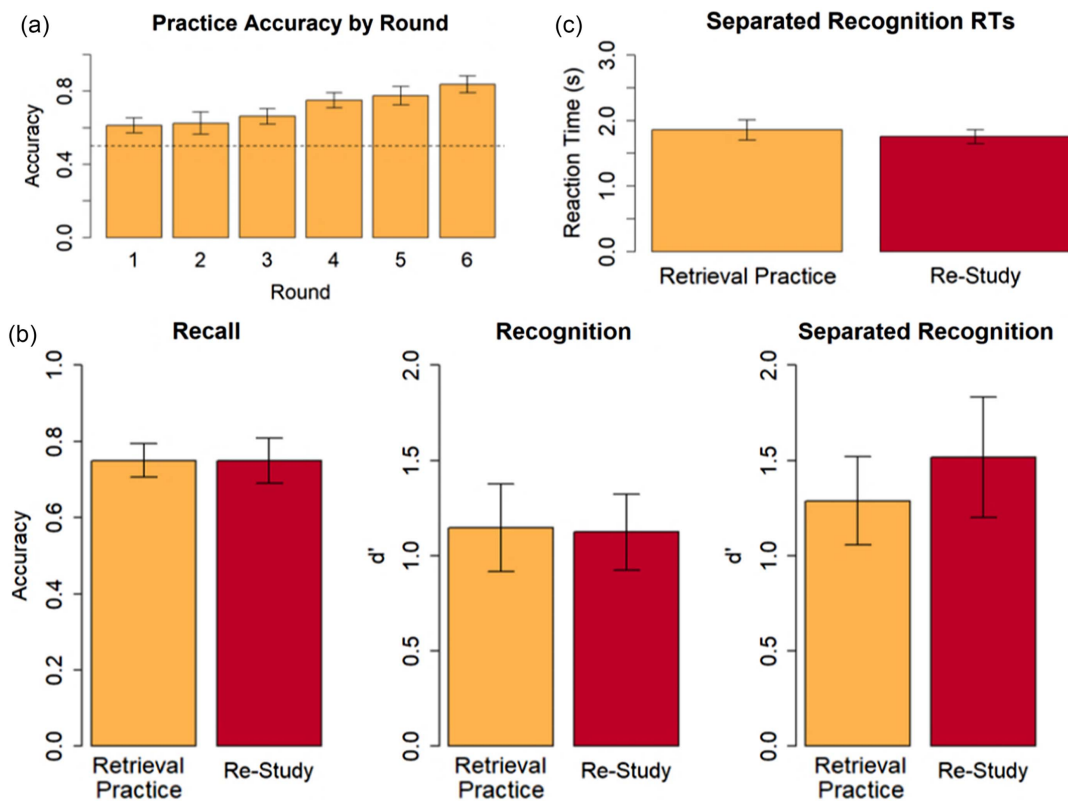
While adding a multiday delay did numerically decrease the performance on the final tests (see the performance in Experiment 1,

Figure 4), it failed to elicit a testing effect. This contrasts with prior findings that adding a multiday delay leads to a stronger testing effect.

Experiment 3

Experiments 1 and 2 presented participants with the whole target object as a form of feedback in the test practice condition. This was done because viewing integrated whole objects may be fundamentally different than viewing separated parts. If the feedback was only in terms of separated parts, then the test practice condition might be at a disadvantage by virtue of giving participants less opportunity to view the whole target objects. However, there is no guarantee that participants paid attention to the visual appearance of the target object when presented for feedback, aside from noting whether their choice was correct. To address this possibility, Experiment 3 presented whole objects not just for the feedback, but also for the choice pairs during test practice. This should force participants to pay attention to the visual appearance of the whole objects in the test practice condition. In addition, one concern with Experiment 1 was that the failure to find differences may have reflected a ceiling effect. This was partially addressed in Experiment 2 through the delay manipulation. Because Experiment 3 did not use a multiday delay,

Figure 5
Results of Experiment 2



Note. (a) Practice accuracy by round of additional familiarization for Experiment 2. The dashed line represents chance-level performance. (b) Recall accuracy, recognition d' , and separated recognition d' for Experiment 2. There was no significant difference between practice and restudy in any of the tests. (c) Reaction times in the separated recognition test for Experiment 2. There was no significant difference between practice and restudy. All error bars represent standard error of the mean. RTs = reaction times. See the online article for the color version of this figure.

the performance was kept below ceiling by reducing the number of rounds of additional familiarization from six to two, and by reducing the amount of time spent restudying or visualizing to 3 s rather than 5 s, as used in Experiments 1 and 2.

Participants

Twenty people participated (18 female, two male) whose ages ranged from 18 to 22 years ($M = 19.7$ years).

Method

The methods for Experiment 3 were identical to those of Experiment 1 except for the following changes. First, only two rounds of additional familiarization occurred, rather than six. Second, the duration of the restudy and the duration of the cue portion of practice were reduced from 5 to 3 s. Third, in the practice and recall trials, the participants now had whole objects to pick between instead of parts alone. The updated practice/recall trial structure is shown in Figure 6.

Results

Participants performed above chance on the practice task (Figure 7a), although their improvement between the first and second rounds was small and failed to reach significance, $t(19) = 0.17, p = .434$. This lack of significant improvement is unsurprising since there were only two rounds of practice in this study. As seen in Figure 7b, restudied objects were remembered significantly better than practiced objects in the final recall test, $t(19) = 2.50, p = .022$. There was no significant difference between practiced objects and restudied objects in the recognition test, $t(19) = 2.00, p = .060$, and separated recognition test, $t(19) = 0.40, p = .694$. In addition to accuracy and d' , the reaction times in the separated recognition test

were compared (Figure 7c). For the reaction time comparison, one participant could not be included because they had no hits in the practice condition (i.e., there were missing data for one participant). Separated recognition reaction times for practiced objects were not significantly faster than the reaction times for restudied objects, $t(18) = 1.87, p = .078$. Plots of individual participant performance can be found in Supplemental Figure S3.

Discussion

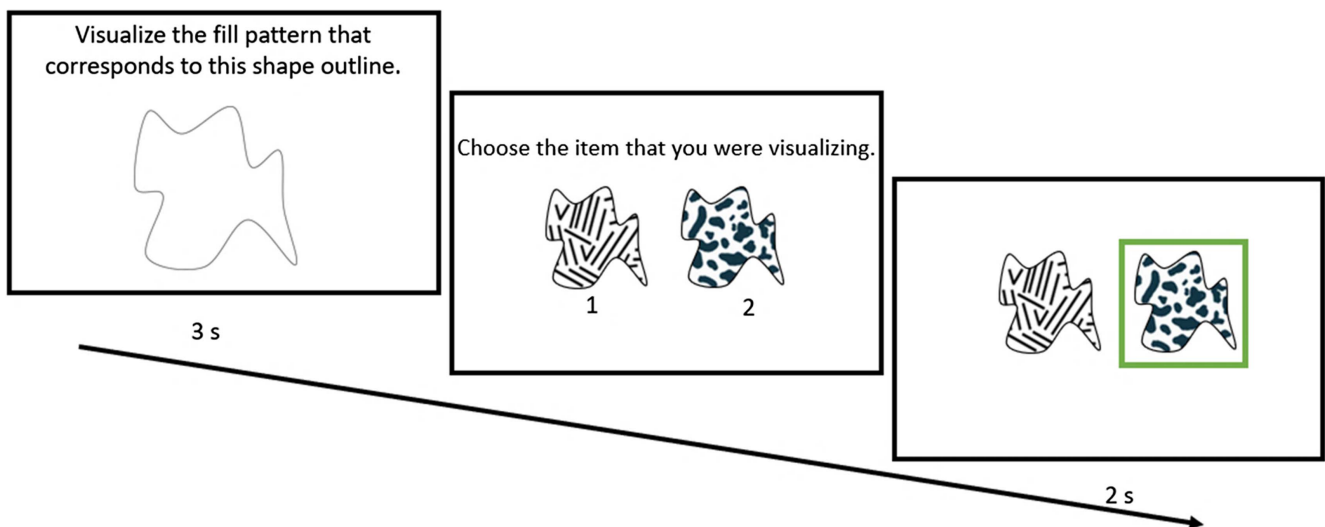
Despite reducing the number and duration of additional familiarization rounds and adding more whole object exposure to the practice task, none of the accuracy or d' measures show a practice benefit. Interestingly, the cued recall final test shows a significant restudy benefit, the opposite of what the testing effect would predict. This restudy benefit may relate to a key difference between Experiment 3 versus Experiments 1 and 2. In Experiments 1 and 2, the retrieval cue for the visualization task was not onscreen at the point when participants made their forced-choice judgments. Thus, participants needed to pay attention during visualization such that, at a minimum, they remembered the retrieval cue. In contrast, Experiment 3 presented the intact whole objects for the forced choice decision, which obviated the need to remember the retrieval cue. In other words, as in the Ferreira and Wimber (2023) study, there was no check to make sure that participants paid attention and attempted visual recall during the visualization task, and this may have reduced any benefits of retrieval practice.

Experiment 4

Experiment 3 attempted to equate restudy and test practice by ensuring that *both* involved ample opportunity to view the target objects as integrated whole objects rather than separated parts.

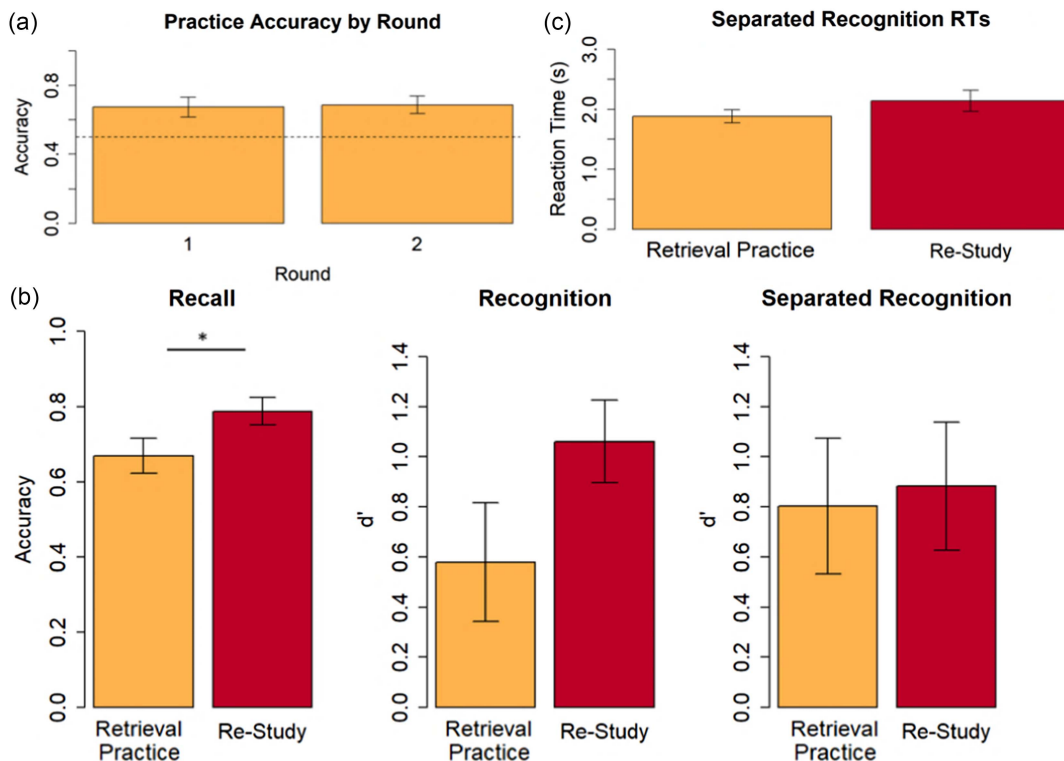
Figure 6

Updated Practice and Recall Task Structure for Experiment 3



Note. Participants picked between the whole objects instead of the individual parts. The cue could be either the shape outline (as shown here) or the fill pattern. A green box provided feedback following retrieval practice trials but, as in all experiments, feedback was not presented during the final recall test. See the online article for the color version of this figure.

Figure 7
Results of Experiment 3



Note. (a) Practice accuracy by round of additional familiarization for Experiment 3. The dashed line represents chance-level performance. (b) Recall accuracy, recognition d' , and separated recognition d' for Experiment 3. Practice recall accuracy was significantly worse than restudy recall accuracy. (c) Reaction times in the separated recognition test for Experiment 3. Reaction times for practiced objects were not significantly lower than reaction times for restudied objects. All error bars represent standard error of the mean. RTs = reaction times. See the online article for the color version of this figure.

* $p < .05$.

Experiment 4 was a second attempt at equating the two conditions by ensuring that *neither* restudy nor test practice involved viewing integrated whole objects. In addition, there is a further concern with Experiments 1–3 that we attempted to correct in Experiment 4. It is known that the benefits of test practice can be tempered if people consistently give wrong answers and learn those wrong answers; indeed, the magnitude of the test practice advantage is related to the accuracy of test practice (Jang et al., 2012; Kornell et al., 2011). Effectively, when people give wrong answers, the familiarity of those wrong answers increases, which can produce interference on the final test. In Experiments 1–3, the part lures (Experiments 1 and 2) or whole object lures (Experiment 3) used during test practice were identical to the lures that were used on the final recognition test for the test practice families. This may have made those lures more familiar, which would work against good performance in the test practice condition (but not in the restudy condition because the final test lures for this condition were never seen prior to the final test). Therefore, a second goal of Experiment 4 was to equate the restudy and test practice conditions in terms of advance exposure to the wrong choice options, or lures, that would appear on the final test. This should equate the conditions in terms of familiarity for the choice options. To equate lure exposure in Experiment 4, in both test practice *and* restudy conditions, both the correct option and the

incorrect (lure) option appeared on the screen during additional familiarization (i.e., during practice retrieval or restudy). The only difference between restudy and test practice was that in test practice, participants needed to select an option before viewing a feedback box that indicated the correct choice, whereas this feedback box was immediately provided in the restudy condition without any choice needed.

Participants

Twenty-four people participated, two of whom were excluded due to the experiment missing the filler task. Of the remaining 22 participants (18 female, three male, one nonbinary), ages ranged from 18 to 23 years ($M = 19.6$ years).

Method

The methods for Experiment 4 were identical to those of Experiment 1 except that no whole object feedback was used, and the restudy trials were changed to better approximate the practice trials by removing whole object exposure along with other minor alterations. The restudy trials no longer employed the likeability rating task; instead, restudy trials used a task identical to that of the

practice trials except that all of the information was shown on one screen simultaneously, and the correct answer was highlighted (see Figure 1), with no response required. In contrast, in the practice trials, first the cue was presented, and next the choice options appeared on a subsequent screen. The key difference between practice and restudy trials was that in practice trials, participants had to select the correct part, whereas in restudy trials, they were directly shown the correct part with a highlighting box. The restudy trial duration was increased to 6 s so that it better approximated the total practice trial duration. A sample restudy trial is shown in Figure 8. Most importantly, since whole object feedback was removed from the practice trials, and the restudy trials were redesigned, no whole object exposure was given in either practice or restudy.

Results

Participants' performance appeared to improve, numerically, on the practice task across the rounds of additional familiarization (Figure 9a), although it did not improve significantly across blocks per a linear analysis, $t(21) = 0.71, p = .242$. This lack of a significant linear improvement is likely due to an anomalous spike in the performance in Round 2. As seen in Figure 9b, restudied objects were remembered significantly better than practiced objects in recall, $t(21) = 3.53, p = .002$; recognition, $t(21) = 2.25, p = .035$; and separated recognition, $t(21) = 2.75, p = .012$. In addition to accuracy and d' , the reaction times in the separated recognition test were compared (Figure 9c). There was no significant difference between reaction times for practiced objects and restudied objects, $t(21) = 0.15, p = .884$. Plots of individual participant performance can be found in Supplemental Figure S4.

Discussion

Surprisingly, even though restudy and test practice were nearly identical in terms of exposure to target and lure parts and no exposure to whole objects, the performance was significantly worse for practiced items than for restudied items, as seen in the accuracy and d' for all three of the final tests. This suggests that the benefit

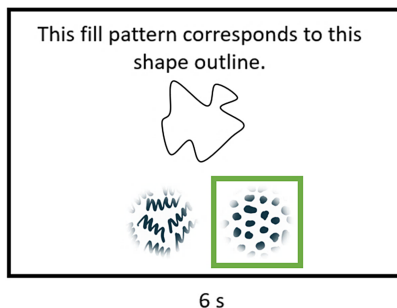
of restudy is related to being told the correct answer without needing to make a choice first—something that is in some sense the exact opposite of the typical explanation of the more commonly found test practice benefit.

Experiment 5

In Experiment 4, above, we addressed the concern that exposure to the wrong choice options, in the practice condition but not the restudy condition, may have caused interference in the practice condition by increasing familiarity for the incorrect “lure” options at the final test. However, in several prior studies, including Experiment 4, there is a further source of interference that applies only to the practice condition—response interference. In Experiments 1–4, for two stimulus families, both targets of the family were assigned to the test practice condition (see the left side of Figure 10), whereas for the other two families, both targets were assigned to restudy. In addition, the lures used in the final 2AFC tests (i.e., the “wrong” part or the “wrong” object) were always drawn from the same family as the target object. This meant that, for each “test practice” stimulus family, participants were given repeated exposure during test practice to all four choice options for that family (e.g., two outlines and two fill patterns for Experiments 1, 2, and 4, and all four objects for Experiment 3). Moreover, in Experiments 1, 2, and 4, in which stimulus *parts* served as the two choice options, all options served as both target (the correct option) and lure (incorrect option) across different practice test trials, depending on which cue was provided. This potentially led to response interference in the practice condition. That is, for practiced items, both options on every final cued recall test were associated with both response options (“target” and “lure”). In contrast, for restudied items, the two options on final cued recall tests either had never been seen in isolation (Experiments 1 and 2) or had never been associated with any response (Experiments 1, 2, and 4).

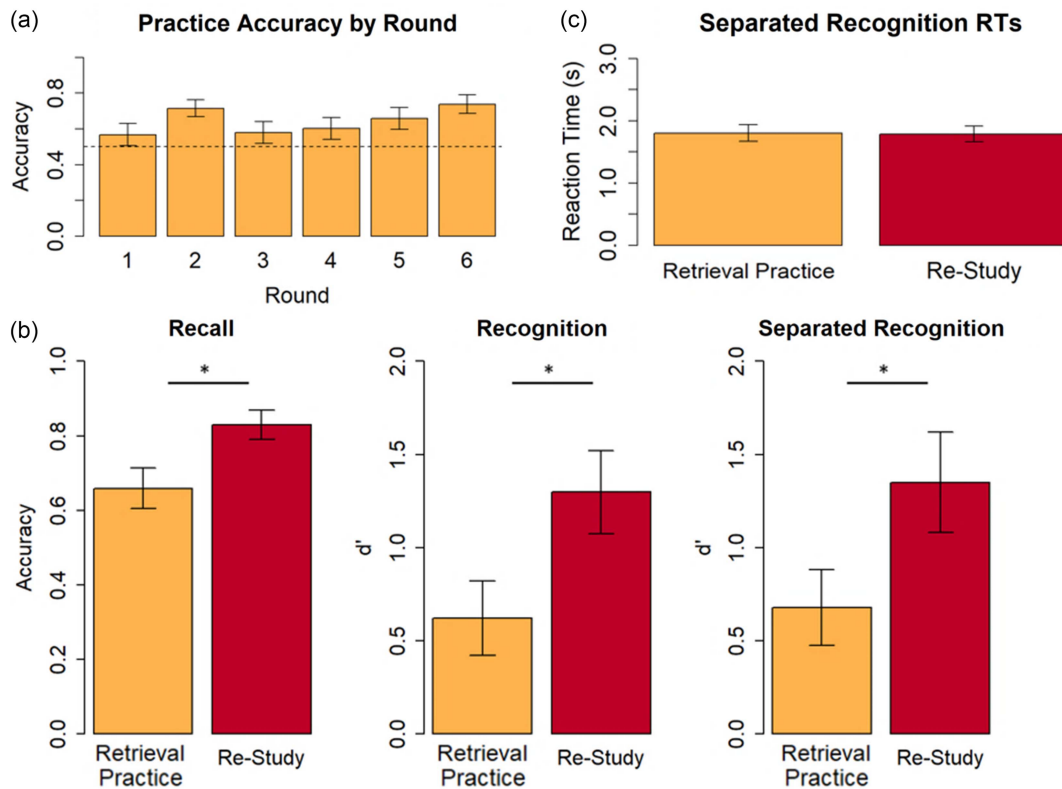
Experiment 5 changed the design such that one target object from a family was assigned to the restudy condition and the other target object from that same family was assigned to the test practice condition. Note that lures for retrieval practice were always selected from the same family as the target object. Therefore, for the practice tests of Experiment 5, when a particular pair of choice alternatives appeared (e.g., a choice between a “herringbone” fill pattern with alternating slanted lines and a fill pattern with diagonally slanted oval blobs; see the bottom-right family in Figure 10), the same choice option was the correct choice for any practice test that contained these two choices, regardless of the retrieval cue, because these choices only ever appeared with one retrieval cue (e.g., for this choice, the correct answer would always be the diagonally slanted oval blobs). In contrast, for Experiments 1, 2, and 4 during retrieval practice, sometimes one choice option was correct while other times the other option was correct, depending on the just-presented retrieval cue (e.g., for the toothlike outline retrieval cue, the herringbone fill pattern was correct, but for the kite-shaped outline retrieval cue, the diagonally slanted oval blobs fill pattern was correct). Thus, for Experiment 5, there was no longer response interference for the retrieval practice condition because there was a consistent correct answer during retrieval practice. Furthermore, this introduced a sort of response interference to the restudy condition because there was a consistent correct choice during retrieval practice for a particular pair of choice options, but that choice was

Figure 8
Sample Restudy Trial for Experiment 4



Note. Instead of seeing the whole object, participants saw a cue along with two choice options, and the correct corresponding part was highlighted. The cue could be either the shape outline (as shown here) or the fill pattern. The choice options were the two instances of the complementary part: fill pattern (as shown here) or shape outline. See the online article for the color version of this figure.

Figure 9
Results of Experiment 4



Note. (a) Practice accuracy by round of additional familiarization for Experiment 4. The dashed line represents chance-level performance. (b) Recall accuracy, recognition d' , and separated recognition d' for Experiment 4. Performance for practiced objects was significantly worse than performance for restudied objects in all three tests. (c) Reaction times in the separated recognition test for Experiment 4. There was no significant difference between practice and restudy. All error bars represent standard error of the mean. RTs = reaction times. See the online article for the color version of this figure.

* $p < .05$.

the wrong choice on the final test whenever the presented retrieval cue was the one from the restudy condition.

Participants

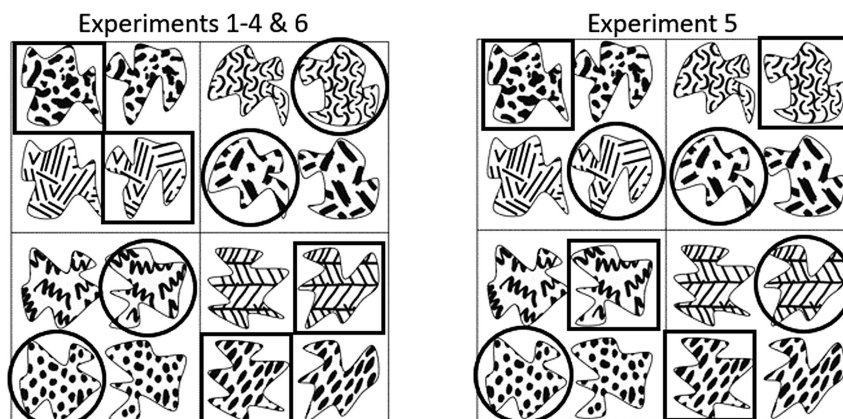
Twenty-seven people participated (23 female, four male) whose ages ranged from 18 to 21 years ($M = 20.0$ years).

Method

The procedure for Experiment 5 was identical to Experiment 1 except that the method of object assignment to conditions was changed, no whole object feedback was given in the practice condition, and there were 10 rounds of additional familiarization. We modified the object assignment because of the possible confound created by response interference. In Experiments 1, 2, and 4, for practiced items, both part options on every final 2AFC cued recall test (the target and the lure) had served as both target and lure during practice retrieval trials. For instance, Fill Pattern A would have been a target when practicing retrieval of Object 1, but a lure when practicing retrieval of Object 2. This likely created competing associations with the category of "target" (correct) versus the category of "lure"

(incorrect) for both fill pattern options. No such interference existed for the restudied items in Experiments 1, 2, and 4 because the two options on the final cued recall tests either had never been seen in isolation or had never been associated with any response. This interference, which potentially influenced practiced items but not restudied items, may have masked any testing effect. This possibility is addressed with Experiment 5.

In Experiment 5, instead of assigning both target objects from a given family to the same condition, one target object from each family was assigned to restudy and the other target object from that family was assigned to practice. Sample object assignments are shown in Figure 10. Note that the two parts used in the final 2AFC tests (e.g., the two outline shapes or the two fill patterns) were always drawn from the same family of four items as the target object. The new object assignment method equated the part exposure for practiced and restudied items: With the new method, for *both* item types, the parts that appeared as the two options on the final recall test (i.e., the outline shapes or fill patterns) had been seen previously during practice retrieval. Furthermore, because only one object per family underwent practice retrieval, only one of the two options for each part (i.e., only one fill pattern and only one shape) ever served as the "correct" part, preventing the formation of competing associations

Figure 10*Overview of Changes Made to Object Assignment for Experiment 5*

Note. A sample assignment of objects to conditions is shown here. Objects encased in a square were targets assigned to the practice condition, objects encased in a circle were targets assigned to the restudy condition, and objects with no shape encasing them were used as lures on the final test.

(i.e., associations of the same part with both target status and lure status). This removed the source of interference that may have affected prior studies. In fact, this should “stack the deck” in favor of an advantage for the test practice condition because there was a consistently correct part to choose during test practice when given a particular pair of choice options, and this correct part was also the correct choice on the final cued recall test for the test practice condition. In contrast, for the restudy condition, these same two choice options appeared on the final test, but in this condition, the part that had been the consistently correct option during test practice was now the *incorrect* part.

Results

Participants’ performance steadily improved on the practice task during each subsequent round of additional familiarization (Figure 11a) with a significant trend according to a linear analysis, $t(26) = 6.18, p < .001$. This indicates that participants were learning the objects during retrieval practice. However, as seen in Figure 11b, there was no significant difference between restudied objects and practice objects in the recall test, $t(26) = 1.84, p = .076$. Restudied objects were remembered significantly better than practiced objects in recognition, $t(26) = 4.73, p < .001$, and separated recognition, $t(26) = 5.81, p < .001$. The reaction times in the separated recognition test were compared across conditions (Figure 11c). There was no significant difference between reaction times for practiced objects and reaction times for restudied objects, $t(26) = 1.44, p = .161$. Plots of individual participant performance can be found in Supplemental Figure S5.

Discussion

Unexpectedly, even with equating the exposure to stimulus parts across the practice and restudy conditions and eliminating response interference for the test practice condition, the performance for the objects in the practice condition was numerically worse than in the restudy condition across all of the final tests. This is particularly

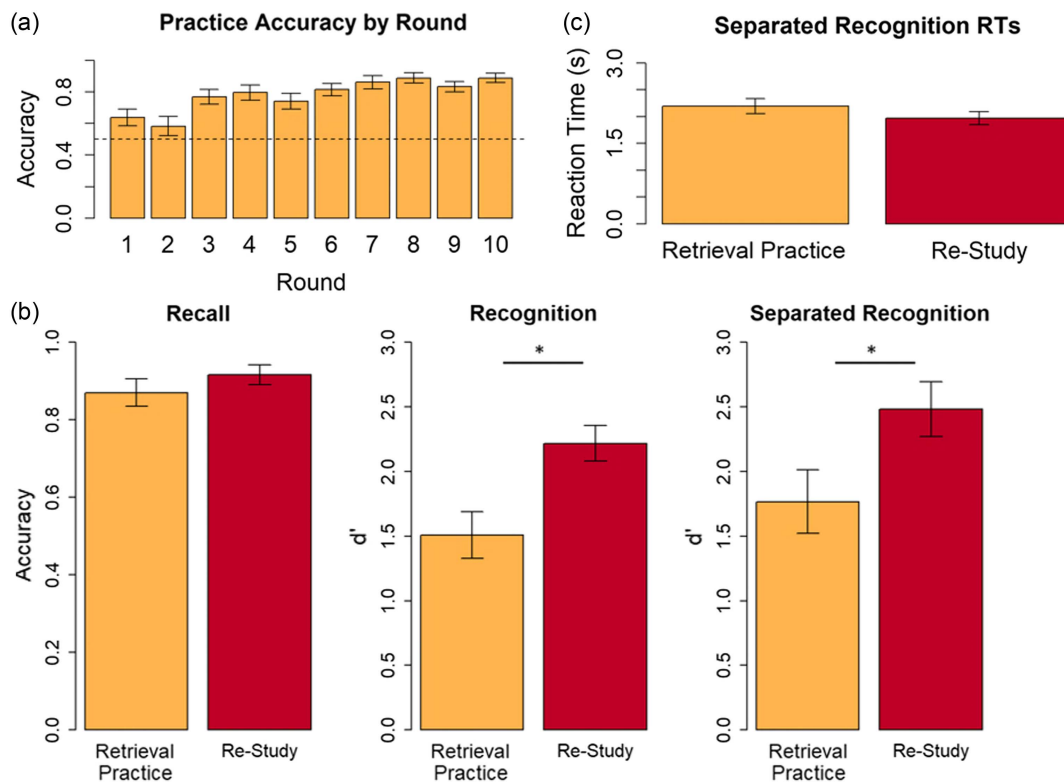
surprising because the modification in this experiment should in fact unfairly bias the results in favor of a testing effect. That is, in Experiment 5, for all objects, the two options (i.e., two fill patterns or two outline shapes) presented on the final 2AFC cued recall test had been presented during practice trials, and, importantly, it was always the same option of the two that was “correct.” Thus, participants could simply learn an association of “good” with one part option and “bad” with the other to perform well during practice retrieval. If participants exploited this rule on the final test, it would produce higher accuracy for practiced objects, without the need for any knowledge of the part-to-part association that was supposed to underlie performance. Furthermore, it would artificially depress the performance on restudied objects because the correct option for restudied objects would always be the “bad” part option from retrieval practice. Thus, even though Experiment 5 provided a strategy that could unfairly inflate accuracy for practiced objects over that for restudied objects, we still did not find a testing effect.

One possible explanation for the failure to find a testing effect despite biasing the results in favor of one is that participants did not learn the association between the retrieval cue and the correct choice during test practice; this is a likely scenario because there was a consistent correct answer to the forced choice, regardless of the retrieval cue. Thus, participants did not need to pay attention to the retrieval cue during the visualization task to do well on the forced choice. This concern also exists for the design of Ferreira and Wimber (2023), which did not require an overt response to the visualization task. In contrast, for Experiments 1, 2, and 4 of the present study, if participants did not pay attention when the retrieval cue was presented, the performance on the practice tests would necessarily be at chance.

Experiment 6

Experiment 4 equated the conditions by presenting the isolated parts for both the restudy and test practice conditions during additional familiarization. Experiment 5 additionally eliminated response interference for the retrieval practice condition considering

Figure 11
Results of Experiment 5



Note. (a) Practice accuracy by round of additional familiarization for Experiment 5. The dashed line represents chance-level performance. (b) Recall accuracy, recognition d' , and separated recognition d' for Experiment 5. Performance was significantly worse for practiced objects than restudied objects in the recognition and separated recognition tests. (c) Reaction times in the separated recognition test for Experiment 5. There was no significant difference between practiced objects and restudied objects. All error bars represent standard error of the mean. RTs = reaction times. See the online article for the color version of this figure.

* $p < .05$.

that there was a consistent correct choice for a given pair of shape/fill attributes on the forced choice test. Furthermore, Experiment 5 introduced response interference to the restudy condition considering that the previously correct choice for a given pair of shape/fill attributes was the incorrect choice if instead the restudy retrieval cue was shown. Nevertheless, restudy resulted in numerically better final test recall as compared to the retrieval practice condition. Experiment 6 goes to the other extreme in an attempt to equate the conditions by ensuring that neither the restudy condition nor the test practice condition entailed advance onscreen presentations of the parts as choice options. This was achieved by having participants draw the corresponding part in the practice task rather than make a 2AFC decision. In addition to equating the conditions in terms of preexposure to the response options, drawing is presumably more effortful and therefore may result in stronger learning. Prior studies have successfully used drawing as an encoding strategy (i.e., Peynircioğlu, 1989; Wammes et al., 2018). This drawing task is also an overt retrieval task, unlike the covert retrieval that we attempted to induce during imagination of the answer in Experiments 1–5. The drawing task ensures that participants are indeed attempting to retrieve when instructed to. Thus, not only

were they forced to pay attention to the retrieval cue during test practice, but they also needed to attempt visual recall because they had to conjure the missing attribute “out of thin air” to guide their drawing.

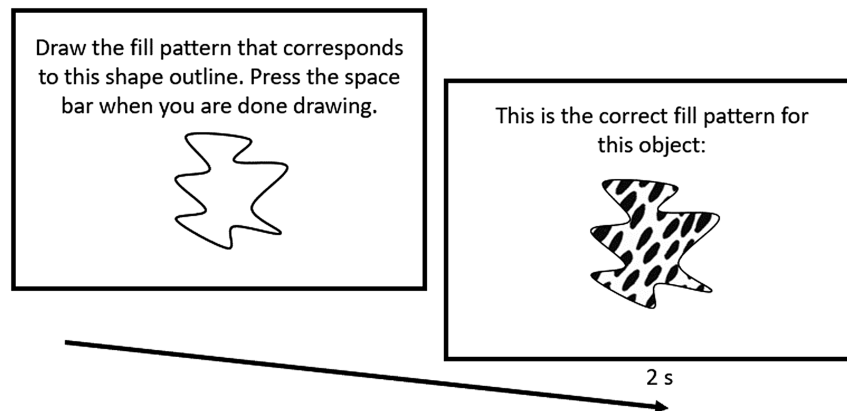
Participants

Twenty-five people participated (20 female, four male, one no response) whose ages ranged from 18 to 23 years ($M = 19.7$ years).

Method

The methods for Experiment 6 match those of Experiment 1 except that the practice trials involved drawing the corresponding part instead of a 2AFC question for the corresponding parts. The modified practice trials can be seen in Figure 12. Participants were shown either the shape or fill pattern of an object and instructed to draw the corresponding part. They were given numbered sheets of paper and a pen for their drawings. The participant pressed the space bar when they were finished with their drawing, and the correct whole object appeared as feedback.

Figure 12
Sample Practice Trial for Experiment 6



Note. Participants see a cue, draw the corresponding part, and then see the whole correct object as feedback. The cue could be either the shape outline (as shown here) or the fill pattern.

Results

For the practice task, participants drew the corresponding part. Three raters, blind to the experimental hypotheses and the correct answer, were presented with each drawing and images of the two parts (either two outlines or two fill patterns) corresponding to that family. The raters picked which of the two options the drawing better matched. The agreement between the raters was moderate to substantial based on Cohen's κ (1 and 2 = 0.528, 2 and 3 = 0.447, 1 and 3 = 0.626). Drawing accuracy as measured by the raters is shown in Figure 13a. Overall, participant drawings were quite poor, but they did significantly improve across blocks per a linear analysis, $t(24) = 1.89, p = .035$. As seen in Figure 13b, restudied objects were remembered significantly better in the recall, $t(24) = 3.48, p = .002$; recognition, $t(24) = 5.57, p < .001$; and separated recognition, $t(24) = 3.91, p < .001$, tests. In addition to accuracy and d' , the reaction times in the separated recognition test were compared (Figure 13c). There was no significant difference between the reaction times for practiced objects and the reaction times for restudied objects, $t(24) = 0.62, p = .542$. Plots of individual participant performance can be found in Supplemental Figure S6.

Discussion

Despite using a more challenging practice task that requires effortful, overt retrieval and despite equating the conditions by ensuring that the choice options did not appear during additional familiarization for either condition, there was still no evidence of a testing effect in Experiment 6.

General Discussion

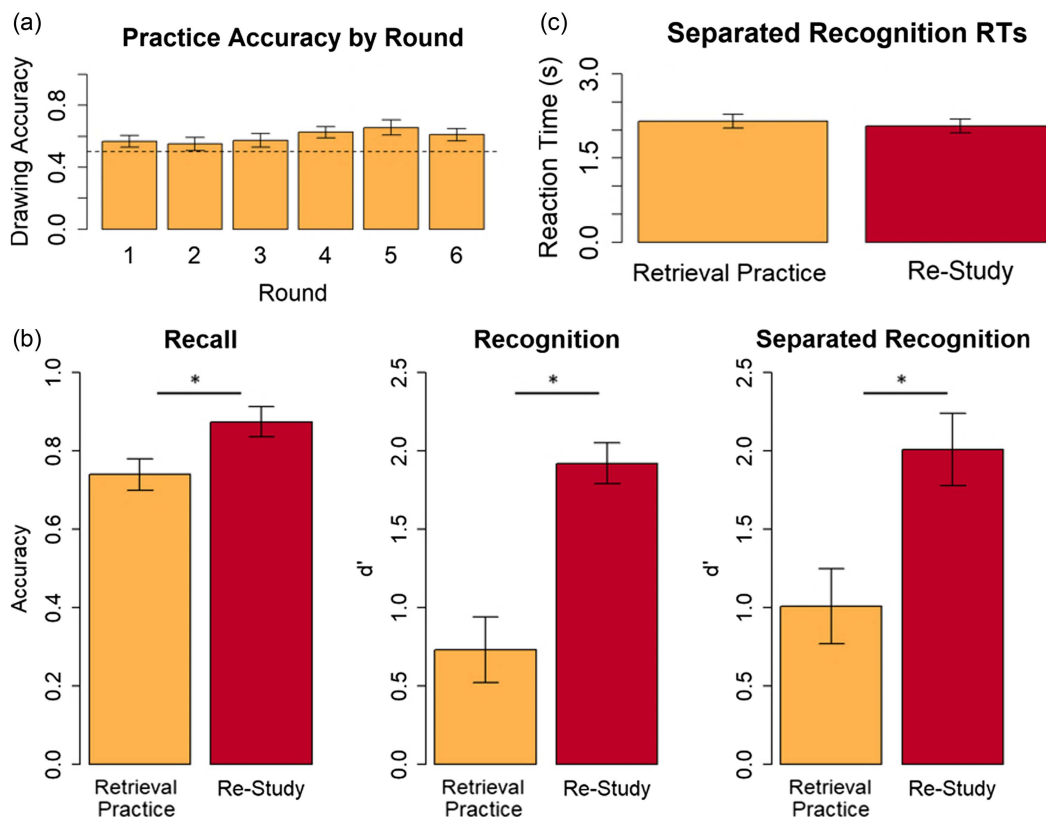
In a series of six experiments, we found no evidence for a purely visual testing effect. In all experiments, memory for objects learned through restudy was either equivalent to or better than memory for objects learned through visual recall practice. Several of the experiments produced clear evidence of learning from test practice (i.e., we observed improvement across rounds of practice), but this learning was never found to be better than the learning from restudy.

This repeated failure to find any benefits for test practice over restudy differs from the large literature of testing effects using meaningful/semantic material. One possible conclusion is that the testing effect does not exist for purely visual abstract material (see Ferreira & Wimber, 2023). However, besides differing from prior studies along the semantic/asemantic dimension, our experiments also differed from the typical testing effect study in at least two other ways: We asked participants to learn associations between two features of the stimuli where those features were themselves highly novel (e.g., between a novel outline shape and a novel fill pattern); and we tested part-to-part associations within an item (i.e., within an object) rather than item-to-item associations (e.g., word-to-word or object-to-object). Finally, it is possible that our adoption of experimental procedures from the verbal/semantic testing effect literature may simply not have been well-suited to visual learning. Below, we discuss each of these possible explanations in greater detail.

First, it is possible that the testing effect exists only for stimuli with semantic content, for example, because testing enhances retrieval by strengthening the semantic network of the to-be-retrieved memories. This explanation is in line with theories advocating that the testing effect is semantic content-mediated, including the elaborative retrieval hypothesis (Carpenter, 2009), the mediator effectiveness hypothesis (Pyc & Rawson, 2010), and the fuzzy trace theory (Bouwmeester & Verkoeijen, 2011; Reyna & Brainerd, 1995). In the present experiments, because the stimuli were abstract, meaningless objects, these theories would predict no benefit of testing on subsequent retrieval, in line with what we found. This explanation diverges from the competing class of theories claiming that the testing effect confers universal retrieval benefits (via transfer-appropriate processing, desirable difficulty, or episodic associations with context; Bjork & Bjork, 2011; Hopper & Huber, 2018; Karpicke et al., 2014; C. D. Morris et al., 1977).

Some prior evidence supports a semantic content-mediated account: Lifanov et al. (2021) found that repeated recall of word–image pairs, compared to restudy, increased the reaction time retrieval advantage for conceptual versus perceptual features of the memory. In other words, the testing effect appeared to selectively enhance the speed of retrieving conceptual information. Indeed, there is evidence

Figure 13
Results of Experiment 6



Note. (a) Practice accuracy by round of additional familiarization as determined by drawing ratings for Experiment 6. The dashed line represents chance-level performance. (b) Recall accuracy, recognition d' , and separated recognition d' for Experiment 6. Performance for practiced objects was significantly worse than performance for restudied objects in all three tests. (c) Reaction times in the separated recognition test for Experiment 6. There was no significant difference between practiced objects and restudied objects. All error bars represent standard error of the mean. RTs = reaction times. See the online article for the color version of this figure.

* $p < .05$.

that not just the testing effect but visual memory in general may be dependent on semantic content: Although visual long-term memory can store thousands of everyday objects with a high degree of detail (Brady et al., 2008), visual long-term memory is quite poor for meaningless visual items encoded under similar conditions, that is, one brief presentation per image (Shoval et al., 2023). Along these lines, Ferreira and Wimber (2023) found retrieval practice benefits for meaningful word–image pairs, but not for paired associates with less semantic content (word–squiggle pairs). However, as noted in the introduction, their design may have biased the results against a testing effect for the meaningless visual objects because the memory task involved the association of the object with a word—by focusing the task on this association, participants may have used semantic associations to guide their responses (e.g., exploiting what the squiggle reminded them of), but such a strategy would of course work better for a word paired with a picture of an everyday object than a squiggle. In contrast, the present study asked participants to associate two different visual aspects of the meaningless object rather than relating the object to a word, and yet there was no testing effect for meaningless visual objects in any of the six experiments. In

opposition to these failures to find a testing effect for meaningless visual objects, Gates (1917) found a testing effect for nonsense syllables, which have no semantic content (although it is possible that participants associated them with real words as a mnemonic strategy), and Boutin et al. (2013) found a testing effect for motor movements. In summary, the current results could be taken as additional evidence supporting the claim that the testing effect is mediated by semantic content, but we cannot make this claim definitively because we did not manipulate semantic content. Next, we consider some alternative explanations of these results.

A second explanation compatible with our findings focuses on our use of highly novel stimuli. The stimuli in these experiments were simple schematic outlines and fill patterns that were devoid of color, three-dimensional form, shading, and reflectance. It may be that such two-dimensional schematic drawings are not easily represented by the visual system unless there is a great deal of prior experience with the corresponding objects in the real world. Consider, for instance, that the initial learning in our study may have been very poor for these novel stimuli (e.g., the results of Experiment 6 indicate that participants struggled to draw the object parts after initial learning).

One theory that explicitly predicts a weak or absent testing effect under such conditions is the “online consolidation” account of retrieval-mediated learning (Antony et al., 2017). This account makes similar predictions to semantic content-mediated accounts but differs by proposing that retrieval serves to consolidate new memories and shift their dependency from the hippocampus to the neocortex. This theory predicts that retrieval practice should most effectively enhance retention when learning can capitalize on preexisting knowledge schemas in the neocortex, suggesting that the testing effect should be reduced for novel materials without prelearned representations, for example, nonsense syllables or meaningless visual images. On this account, the failure to find a testing effect in our experiments arose from the use of highly novel, unfamiliar stimuli rather than the use of meaningless stimuli. It may be difficult to learn to associate two things if you possess poorly established representations of those two things, regardless of the modality of the representations. Applying this logic to a verbal testing effect paradigm, an experiment asking participants to learn to associate two nonpronounceable nonwords might likewise fail to produce a testing effect (the nonwords would need to be nonpronounceable to make sure they are truly novel).

A third possible explanation of our results concerns the requirement to learn within-item, part-to-part associations that depend on relatively low-level visual representations. Typical testing effect studies employ item-to-item, paired-associate learning. It is therefore possible that the testing effect occurs only for the learning of something that is complex and high dimensional. Highly complex representations—such as associated word pairs, the associated collection of objects that form a scene, or the association of an item with its study context—are assumed by many theories of cognition to depend on the hippocampus (Bussey & Saksida, 2007; Eichenbaum et al., 1992; Sutherland & Rudy, 1989). Such a theory might therefore imply that the testing effect is “hippocampal,” tapping specialized representations or operations that are housed or performed only in the hippocampus. This explanation is different from the “online consolidation” account: Consolidation involves a shift from the hippocampus to the neocortex, whereas a “high-dimensional content” account ties the testing effect to the hippocampus. On this account, the failure to find a testing effect in our experiments arose from the use of part-to-part associations within an object, rather than object-to-object associations, as in a visual scene. The analogous experiment with verbal stimuli would ask participants to learn single pronounceable nonwords through stem-completion practice (i.e., part-to-part learning of items without meaning, since semantics confer high dimensionality), and a “high-dimensional content” account would predict failure to produce a testing effect. If so, this would be a direct falsification of the primary and convergent retrieval model of Hopper and Huber (2018, 2019), which claims that recall practice provides a unique opportunity for learning the associations between the parts within a target item.

Finally, there exists a fourth possible explanation for the present findings. Perhaps the testing effect does indeed exist for abstract, novel, visual stimuli of low complexity, but our experimental paradigm did not use a type of test practice that was well-suited to low-level visual learning. To our knowledge, this was the first attempt to find a purely visual testing effect (the study of Ferreira & Wimber, 2023, comes close to doing so, but used words for retrieval cues), and as a starting point, we employed retrieval practice procedures that produce a robust testing effect with verbal/semantic memories. In our study, participants saw a static, isolated cue (e.g., a fill pattern without any outline shape) and were given several

seconds to retrieve before receiving feedback. This kind of practice differs from real-life visual learning in two important ways. First, in real life, visual recall often occurs through partial occlusion of a whole object rather than through viewing a disembodied fill pattern or a texture-less outline shape. Consider for instance the visual pattern completion process that occurs when viewing an object through a picket fence or when presented with a degraded image (Gorlin et al., 2012; Warrington & Weiskrantz, 1968, 1970). Second, in real life, answer feedback is often presented to the visual system very rapidly rather than after seconds of attempted retrieval. For example, when moving, your view of an object through a picket fence is constantly changing, providing rapid feedback for previously obscured sections of the object. By adopting procedures tailored to verbal material, we may have inadvertently handicapped visual learning.

One possible concern regarding the interpretation of these data is that performance showed a significant linear improvement across retrieval practice trials in only three out of six experiments (Experiments 2, 5, and 6), with no significant linear trend in the other three (Experiments 1, 3, and 4). Could this indicate that our stimuli were simply too difficult to learn, or that we failed to induce sufficient engagement with the task in our participants, such that there was very little learning of any kind during the additional study rounds, either restudy or retrieval practice? There is evidence against this interpretation in Experiments 3 and 4, where although retrieval practice failed to linearly improve across rounds, there was a significant restudy advantage at the final test. In other words, although retrieval practice was relatively ineffective, restudy improved subsequent memory over and above the initial encoding experience. This evidence should be considered alongside the likely reasons noted in the individual experimental discussions for our failure to detect linear trends in Experiments 1, 3, and 4. Together, the data point to the interpretation that there is no testing effect in this paradigm because restudy is simply more effective than retrieval practice at enhancing memory.

In sum, the present failure to find any evidence for a purely visual testing effect—across six experiments, some of which implemented heavy-handed attempts to tip the scale in favor of a retrieval-induced memory enhancement—yields an unexpected puzzle. At present, multiple alternative accounts compete to explain this puzzle, and further research is needed to adjudicate between those accounts. Regardless of the explanation, this series of results is quite unexpected given the robust nature of the testing effect and draws attention to several avenues for investigation that could help shed much-needed light on the mechanisms, and the limits, of the testing effect.

References

- Abbott, E. E. (1909). On the analysis of the factor of recall in the learning process. *The Psychological Review: Monograph Supplements*, 11(1), 159–177. <https://doi.org/10.1037/h0093018>
- Aenugu, S., & Huber, D. E. (2021). Asymmetric weights and retrieval practice in an autoassociative neural network model of paired-associate learning. *Neural Computation*, 33(12), 3351–3360. https://doi.org/10.1162/neco_a_01444
- Antony, J. W., Ferreira, C. S., Norman, K. A., & Wimber, M. (2017). Retrieval as a fast route to memory consolidation. *Trends in Cognitive Sciences*, 21(8), 573–576. <https://doi.org/10.1016/j.tics.2017.05.001>
- Bjork, E. L., & Bjork, R. A. (2011). Making things hard on yourself, but in a good way: Creating desirable difficulties to enhance learning. In M. A.

- Gernsbacher, R. W., Pew, L. M., Hough, & J. R. Pomerantz (Eds.), *Psychology and the real world: Essays illustrating fundamental contributions to society* (pp. 56–64). Worth Publishers.
- Boutin, A., Panzer, S., & Blandin, Y. (2013). Retrieval practice in motor learning. *Human Movement Science*, 32(6), 1201–1213. <https://doi.org/10.1016/j.humov.2012.10.002>
- Bouwmeester, S., & Verkoeijen, P. P. (2011). Why do some children benefit more from testing than others? Gist trace processing to explain the testing effect. *Journal of Memory and Language*, 65(1), 32–41. <https://doi.org/10.1016/j.jml.2011.02.005>
- 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>
- Bussey, T. J., & Saksida, L. M. (2007). Memory, perception, and the ventral visual-perirhinal-hippocampal stream: Thinking outside of the boxes. *Hippocampus*, 17(9), 898–908. <https://doi.org/10.1002/hipo.20320>
- 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. <https://doi.org/10.1037/a0017021>
- Carpenter, S. K., & DeLosh, E. L. (2005). Application of the testing and spacing effects to name learning. *Applied Cognitive Psychology*, 19(5), 619–636. <https://doi.org/10.1002/acp.1101>
- Carpenter, S. K., & Kelly, J. W. (2012). Tests enhance retention and transfer of spatial learning. *Psychonomic Bulletin & Review*, 19(3), 443–448. <https://doi.org/10.3758/s13423-012-0221-2>
- Carpenter, S. K., & Pashler, H. (2007). Testing beyond words: Using tests to enhance visuospatial map learning. *Psychonomic Bulletin & Review*, 14(3), 474–478. <https://doi.org/10.3758/BF03194092>
- Congleton, A., & Rajaram, S. (2012). The origin of the interaction between learning method and delay in the testing effect: The roles of processing and conceptual retrieval organization. *Memory & Cognition*, 40(4), 528–539. <https://doi.org/10.3758/s13421-011-0168-y>
- Coppens, L. C., Verkoeijen, P. P., & Rikers, R. M. (2011). Learning Adinkra symbols: The effect of testing. *Journal of Cognitive Psychology*, 23(3), 351–357. <https://doi.org/10.1080/20445911.2011.507188>
- Duchaine, B., & Nakayama, K. (2006). The Cambridge face memory test: Results for neurologically intact individuals and an investigation of its validity using inverted face stimuli and prosopagnosic participants. *Neuropsychologia*, 44(4), 576–585. <https://doi.org/10.1016/j.neuropsychologia.2005.07.001>
- Eichenbaum, H., Otto, T., & Cohen, N. J. (1992). The hippocampus—What does it do? *Behavioral and Neural Biology*, 57(1), 2–36. [https://doi.org/10.1016/0163-1047\(92\)90724-I](https://doi.org/10.1016/0163-1047(92)90724-I)
- Ferreira, C. S., & Wimber, M. (2023). The testing effect for visual materials depends on preexisting knowledge. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 49(10), 1557–1571. <https://doi.org/10.1037/xlm0001248>
- Gates, A. I. (1917). *Recitation as a factor in memorizing* (Vol. 40). Science Press.
- Gorlin, S., Meng, M., Sharma, J., Sugihara, H., Sur, M., & Sinha, P. (2012). Imaging prior information in the brain. *Proceedings of the National Academy of Sciences of the United States of America*, 109(20), 7935–7940. <https://doi.org/10.1073/pnas.1111224109>
- Hautus, M. J. (1995). Corrections for extreme proportions and their biasing effects on estimated values of d' . *Behavior Research Methods, Instruments & Computers*, 27(1), 46–51. <https://doi.org/10.3758/BF03203619>
- Hopper, W. J., & Huber, D. E. (2018). Learning to recall: Examining recall latencies to test an intra-item learning theory of testing effects. *Journal of Memory and Language*, 102, 1–15. <https://doi.org/10.1016/j.jml.2018.04.005>
- Hopper, W. J., & Huber, D. E. (2019). Testing the primary and convergent retrieval model of recall: Recall practice produces faster recall success but also faster recall failure. *Memory & Cognition*, 47(4), 816–841. <https://doi.org/10.3758/s13421-019-00903-x>
- Hyde, T. S., & Jenkins, J. J. (1969). Differential effects of incidental tasks on the organization of recall of a list of highly associated words. *Journal of Experimental Psychology*, 82(3), 472–481. <https://doi.org/10.1037/h0028372>
- Jang, Y., Wixted, J. T., Pecher, D., Zeelenberg, R., & Huber, D. E. (2012). Decomposing the interaction between retention interval and study/test practice: The role of retrievability. *Quarterly Journal of Experimental Psychology*, 65(5), 962–975. <https://doi.org/10.1080/17470218.2011.638079>
- Jonker, T. R., Dimsdale-Zucker, H., Ritchey, M., Clarke, A., & Ranganath, C. (2018). Neural reactivation in parietal cortex enhances memory for episodically linked information. *Proceedings of the National Academy of Sciences of the United States of America*, 115(43), 11084–11089. <https://doi.org/10.1073/pnas.1800006115>
- Kang, S. H. (2010). Enhancing visuospatial learning: The benefit of retrieval practice. *Memory & Cognition*, 38(8), 1009–1017. <https://doi.org/10.3758/MC.38.8.1009>
- Karpicke, J. D., Lehman, M., & Aue, W. R. (2014). Retrieval-based learning: An episodic context account. *Psychology of Learning and Motivation*, 61, 237–284. <https://doi.org/10.1016/B978-0-12-800283-4.00007-1>
- Kornell, N., Bjork, R. A., & Garcia, M. A. (2011). Why tests appear to prevent forgetting: A distribution-based bifurcation model. *Journal of Memory and Language*, 65(2), 85–97. <https://doi.org/10.1016/j.jml.2011.04.002>
- Lifanov, J., Linde-Domingo, J., & Wimber, M. (2021). Feature-specific reaction times reveal a semanticisation of memories over time and with repeated remembering. *Nature Communications*, 12(1), Article 3177. <https://doi.org/10.1038/s41467-021-23288-5>
- Macmillan, N. A., & Creelman, C. D. (2004). *Detection theory: A user's guide*. Psychology press. <https://doi.org/10.4324/9781410611147>
- McCarter, A. (2023, June 14). *No evidence of a visual testing effect for novel, meaningless objects*. <https://osf.io/57jsv/>
- Morris, C. D., Bransford, J. D., & Franks, J. J. (1977). Levels of processing versus transfer appropriate processing. *Journal of Verbal Learning and Verbal Behavior*, 16(5), 519–533. [https://doi.org/10.1016/S0022-5371\(77\)80016-9](https://doi.org/10.1016/S0022-5371(77)80016-9)
- Morris, P. E., & Fritz, C. O. (2000). The name game: Using retrieval practice to improve the learning of names. *Journal of Experimental Psychology: Applied*, 6(2), 124–129. <https://doi.org/10.1037/1076-898X.6.2.124>
- Mozer, M. C., Howe, M., & Parshler, H. (2004). Using testing to enhance learning: A comparison of two hypotheses. *Proceedings of the Annual Meeting of the Cognitive Science Society*, 26, 975–980. <https://escholarship.org/uc/item/67r153f3>
- Nairne, J. S., Pandeirada, J. N., & Thompson, S. R. (2008). Adaptive memory: The comparative value of survival processing. *Psychological Science*, 19(2), 176–180. <https://doi.org/10.1111/j.1467-9280.2008.02064.x>
- Peirce, J., Gray, J. R., Simpson, S., MacAskill, M., Höchenberger, R., Sogo, H., Kastman, E., & Lindeløv, J. K. (2019). PsychoPy2: Experiments in behavior made easy. *Behavior Research Methods*, 51(1), 195–203. <https://doi.org/10.3758/s13428-018-01193-y>
- Peynircioğlu, Z. F. (1989). The generation effect with pictures and nonsense figures. *Acta Psychologica*, 70(2), 153–160. [https://doi.org/10.1016/0001-6918\(89\)90018-8](https://doi.org/10.1016/0001-6918(89)90018-8)
- Pyc, M. A., & Rawson, K. A. (2010). Why testing improves memory: Mediator effectiveness hypothesis. *Science*, 330(6002), Article 335. <https://doi.org/10.1126/science.1191465>
- R Core Team. (2021). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, <https://www.R-project.org/>
- Reyna, V. F., & Brainerd, C. J. (1995). Fuzzy-trace theory: An interim synthesis. *Learning and Individual Differences*, 7(1), 1–75. [https://doi.org/10.1016/1041-6080\(95\)90031-4](https://doi.org/10.1016/1041-6080(95)90031-4)

- Roediger, H. L., III, & Karpicke, J. D. (2006a). Test-enhanced learning: Taking memory tests improves long-term retention. *Psychological Science*, 17(3), 249–255. <https://doi.org/10.1111/j.1467-9280.2006.01693.x>
- Roediger, H. L., III, & Karpicke, J. D. (2006b). The power of testing memory: Basic research and implications for educational practice. *Perspectives on Psychological Science*, 1(3), 181–210. <https://doi.org/10.1111/j.1745-6916.2006.00012.x>
- 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>
- Sadil, P., Potter, K. W., Huber, D. E., & Cowell, R. A. (2019). Connecting the dots without top-down knowledge: Evidence for rapidly-learned low-level associations that are independent of object identity. *Journal of Experimental Psychology: General*, 148(6), 1058–1070. <https://doi.org/10.1037/xge0000607>
- Schuetze, B. A., Eglington, L. G., & Kang, S. H. K. (2019). Retrieval practice benefits memory precision. *Memory*, 27(8), 1091–1098. <https://doi.org/10.1080/09658211.2019.1623260>
- Shoval, R., Gronau, N., & Makovski, T. (2023). Massive visual long-term memory is largely dependent on meaning. *Psychonomic Bulletin & Review*, 30(2), 666–675. <https://doi.org/10.3758/s13423-022-02193-y>
- Siler, J., & Benjamin, A. S. (2020). Long-term inference and memory following retrieval practice. *Memory & Cognition*, 48(4), 645–654. <https://doi.org/10.3758/s13421-019-00997-3>
- Sutherland, R. J., & Rudy, J. W. (1989). Configural association theory: The role of the hippocampal formation in learning, memory, and amnesia. *Psychobiology*, 17(2), 129–144. <https://doi.org/10.3758/BF03337828>
- Sutterer, D. W., & Awh, E. (2016). Retrieval practice enhances the accessibility but not the quality of memory. *Psychonomic Bulletin & Review*, 23(3), 831–841. <https://doi.org/10.3758/s13423-015-0937-x>
- Tse, C. S., Balota, D. A., & Roediger, H. L., III. (2010). The benefits and costs of repeated testing on the learning of face-name pairs in healthy older adults. *Psychology and Aging*, 25(4), 833–845. <https://doi.org/10.1037/a0019933>
- Wammes, J. D., Meade, M. E., & Fernandes, M. A. (2018). Creating a recollection-based memory through drawing. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 44(5), 734–751. <https://doi.org/10.1037/xlm0000445>
- Warrington, E. K., & Weiskrantz, L. (1968). New method of testing long-term retention with special reference to amnesic patients. *Nature*, 217(5132), 972–974. <https://doi.org/10.1038/217972a0>
- Warrington, E. K., & Weiskrantz, L. (1970). Amnesic syndrome: Consolidation or retrieval? *Nature*, 228(5272), 628–630. <https://doi.org/10.1038/228628a0>
- Whiffen, J. W., & Karpicke, J. D. (2017). The role of episodic context in retrieval practice effects. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 43(7), 1036–1046. <https://doi.org/10.1037/xlm0000379>

Received April 5, 2024

Revision received August 23, 2024

Accepted October 3, 2024 ■