AMERICAN PSYCHOLOGICAL ASSOCIATION

Journal of Experimental Psychology: General

ISSN: 0096-3445

2024, Vol. 153, No. 7, 1816–1837
https://doi.org/10.1037/xge0001593

The Magnitude of the Testing Effect Is Independent of Retrieval Practice Performance

Jason C. K. Chan[1], Sara D. Davis[2], Aslı Yurtsever[1], and Sarah J. Myers[3]
[1] Department of Psychology, Iowa State University
[2] Department of Psychology, University of North Florida
[3] Department of Psychology, Western Kentucky University

Practicing retrieval is a potent learning enhancer. Theoretical accounts of the testing effect generally suggest that the magnitude of the testing effect is dependent on retrieval practice performance, such that conditions that promote better retrieval practice performance should result in a greater testing effect. Empirical evidence, however, has been mixed. Although some studies showed a positive association between retrieval practice performance and the testing effect, others have shown either no relation or the reverse. In the present study, we experimentally manipulated retrieval practice performance using a retrieval-based response deadline manipulation and an encoding-based study trial manipulation. Across six experiments, the magnitude of the testing effect was independent of retrieval practice performance. However, when we aggregated the data across the experiments, participants with superior retrieval practice performance showed a greater testing effect—an individual difference. This dissociation between experimental and correlational outcomes suggests that the positive relation between retrieval practice performance and the testing effect is not causal, and indeed, simulation data showed that the correlation between retrieval practice performance and testing effect was an artifact. We discuss the challenges these findings present to existing accounts of the testing effect.

Public Significance Statement
Practicing retrieval from memory can enhance learning and retention—the testing effect. However, research has produced mixed evidence about whether the benefits of retrieval depend on practice performance, with some suggesting that conditions that elicit better practice performance produce a greater testing effect, whereas others demonstrating either no effect or the opposite. In the present experiments, we showed that conditions that favored better retrieval practice performance did not produce a greater testing effect. These results challenge existing theories, and they demonstrate that students can reap similar benefits from retrieval practice regardless of whether or not learning conditions are optimized for successful retrieval practice.

Keywords: testing effect, retrieval practice, response deadline, individual difference, learning and memory

Research on the testing effect has experienced substantial growth since the mid-2000s, and the wealth of empirical data has given rise to several meta-analytic reviews of the literature (Adesope et al., 2017; Pan & Rickard, 2018; Rowland, 2014; Yang et al., 2021). Despite this accumulation of knowledge, a question has persisted from the early days of research on the testing effect: What is the relationship between retrieval practice (initial test) performance and the magnitude of the testing effect?

This article was published Online First May 2, 2024.
Ashleigh Maxcey served as action editor.
Jason C. K. Chan https://orcid.org/0000-0003-4221-2979
Some of the data reported in this article were presented in a poster at the 2016 meeting of the Psychonomic Society in Boston, Massachusetts, United States. This work is partly supported by the National Science Foundation Science of Learning and Augmented Intelligence Grant 2017333 to Jason C. K. Chan.

Jason C. K. Chan served as lead for conceptualization, formal analysis, funding acquisition, resources, validation, visualization, writing–original draft, and writing–review and editing and served in a supporting role for data curation. Sara D. Davis served as lead for data curation, investigation, and software, contributed equally to conceptualization and resources, and served in a supporting role for writing–original draft and writing–review and editing. Aslı Yurtsever contributed equally to visualization and served in a supporting role for formal analysis, validation, writing–original draft, and writing–review and editing. Sarah J. Myers served in a supporting role for data curation, investigation, writing–original draft, and writing–review and editing. Jason C. K. Chan and Sara D. Davis contributed equally to methodology, supervision, and project administration.
Correspondence concerning this article should be addressed to Jason C. K. Chan, Department of Psychology, Iowa State University, 1347 Lagomarcino Hall, Ames, IA 50011, United States. Email: ckchan@iastate.edu

A common concern in the testing effect literature, especially in its formative years, was the belief that a testing effect would only manifest when retrieval practice performance was sufficiently high, particularly when no corrective feedback was issued or when restudy was employed as the comparative control condition. The rationale was that when initial test performance was low, retrieval practice would only benefit the small subset of items that were reexposed via recall. In contrast, participants in a typical restudy condition would be reexposed to all of the to-be-learned items (Kornell et al., 2011). Thus, the broader reexposure in the restudy condition might overshadow the benefits of retrieval practice. Indeed, researchers interested in the testing effect have sometimes gone to great lengths to ensure high retrieval practice performance (Kuo & Hirshman, 1996; Thompson et al., 1978; Toppino & Cohen, 2009).

A similar concern exists when comparing participants with greater retrieval practice performance to those with poorer performance, regardless of whether or not the control condition involves restudying. When higher performers recall more studied items during the initial test, they are reexposed to a greater proportion of these items relative to lower performers. Further, if these lower performing participants' recall performance is at floor levels (i.e., they recall nothing or close to nothing) during the initial test, retrieval practice cannot enhance later performance because nothing is initially retrieved. So logically, the magnitude or likelihood of observing a testing effect should be influenced by retrieval practice performance.[1] This logic holds whether the benefits of testing arise from the act of retrieval itself (a direct benefit; Roediger & Karpicke, 2006) or from postretrieval exposure (an indirect benefit). However, such floor-level scenarios are neither compelling nor illuminating because (a) floor effects are difficult to interpret, and (b) from an educational perspective, students are unlikely to engage in retrieval practice if they anticipate floor-level performance and would most likely prioritize restudying first (Rea et al., 2022). Therefore, in the present study, we investigated the relationship between retrieval practice performance and the testing effect while avoiding floor-level retrieval practice performance (operationally defined here as $<30\%$ accuracy).

Theoretical Considerations

Existing explanations of the testing effect generally predict a positive relationship between retrieval practice performance and the testing effect. For example, according to the elaborative retrieval account (Carpenter, 2009), participants might activate semantic associates of the target when attempting retrieval, which can augment the number of retrieval routes to the target (see also Chan, 2009; Chan et al., 2006) and facilitate subsequent retrieval. These processes are believed to be absent when participants do not engage in retrieval. Based on this idea, the more targets one can retrieve during retrieval practice, the more one should benefit from this form of elaboration. Hence, retrieval practice performance and the testing effect should have a positive association.

The episodic context account (Karpicke et al., 2014) posits that every encoding episode is bound to a context. When one initially learns an item, that item is encoded with the study context (e.g., the lecture hall where one encounters the information). When the person later retrieves the item, it is reencoded with a new context (e.g., the dorm room where the person performs retrieval practice). Consequently, the item is now associated with both the initial study context and the retrieval practice context. This episodic context updating increases the likelihood that the learner will be able to retrieve the item again later because the additional context aids the learner in restricting the search set, reducing interference during later subsequent retrieval. Similar to the elaborative retrieval account (Carpenter, 2009), the episodic context account also leads one to predict a positive association between retrieval practice performance and the testing effect, because more items retrieved during the initial test would mean more items benefiting from context updating.

The bifurcation account (Kornell et al., 2011) is a descriptive model. It assumes that memories can be described as a frequency distribution along a memory strength continuum, with stronger memories placed on the right side of the distribution and weaker ones on the left side (Figure 1 in Kornell et al.'s article provides an excellent visualization of the core concepts of the bifurcation account). According to this account, retrieval provides a substantial boost to the memory strength of the subset of retrieved items, causing a bifurcation of the memory strength distribution of the items (i.e., the retrieved items are now much stronger than the nonretrieved items, thereby separating these two sets of items on the strength continuum). In comparison, bifurcation does not occur when learners restudy or are not tested on the to-be learned material. In the case of restudy, memory strength is increased for all items equally, whereas memory strength is unchanged in the absence of an intervention in a no-test condition. The account specifies that all items would suffer some forgetting over time (the distribution shifts leftward), but that a larger portion of items remain retrievable following initial testing due to the bifurcation of the distribution. Applying this logic to the present context, the bifurcation account might also lead one to predict a positive relationship between retrieval practice performance and the testing effect, because more items recalled during the initial test means that more items would move to the right of the strength distribution.

Surprisingly, there is a relative dearth of research that has experimentally tested this assumption in the context of any of these three theoretical frameworks. In the section that follows, we briefly review the literature that provides insights into the observational relationship between retrieval practice performance and the magnitude of the testing effect. To preview, existing evidence spans the gamut, with some studies showing a positive relationship between retrieval practice performance and the testing effect, whereas others show no relationship or even a negative relationship.

Evidence for a Positive Relationship

Data from meta-analyses provide a powerful source of support for the notion that greater retrieval practice performance will likely produce a greater testing effect. For instance, in one of the first comprehensive meta-analytic review of the testing effect literature, Rowland (2014) demonstrated that retrieval practice performance was one of the most powerful moderators of the magnitude of the testing effect. In fact, among studies that did not issue feedback, those with poorer ($\leq.50$) retrieval practice performance produced no testing effect over

[1] Throughout this article, we attempted to avoid using the terms "initial test" and "retrieval practice" interchangeably to reduce confusions. To this end, we used "initial test" when describing the experimental procedure of taking a test, and we used "retrieval practice" when referring to test performance and the general behavioral intervention of using tests to enhance learning.

restudy, whereas those with greater (>.75) retrieval practice performance did. Based on this finding, Rowland argued that "individual studies that show a restudy benefit at short retention intervals may in part reflect low [retrieval practice] performance (and thus lack of reexposure) for test condition items" (p. 18).

In a subsequent meta-analysis investigating the extent to which practicing retrieval would benefit performance on tests that require transfer, Pan and Rickard (2018) showed that retrieval practice performance is a primary predictor of the extent of the testing benefit. Specifically, studies with better retrieval practice performance reported a larger transfer testing effect ($r = .38$). Indeed, retrieval practice performance had such an important influence that it was designated as a key element in Pan and Rickard's three-factor framework. A thorough consideration of this framework is beyond the scope of the present paper because the framework deals with the transfer effects rather than the direct effects of retrieval practice.

Apart from meta-analytic evidence, several experimental studies have also demonstrated a positive relationship. For example, in a classroom study by Cranney et al. (2009), students viewed a video on psychobiology and either took no test, restudied the material, took a test individually, or took a test collaboratively in small groups. After a 1-week retention interval, all students took a final test individually. Participants in the collaborative recall condition ($M = 0.94$) vastly outperformed those in the individual recall condition ($M = 0.47$) during the initial test, and this difference in retrieval practice performance was largely maintained during the final test ($M_{\text{collaborative}} = 0.71$, $M_{\text{individual}} = 0.47$, M_{restudy} and $M_{\text{no-test}} = 0.25$).

In Agarwal et al. (2008), participants completed an open-book or closed-book initial test after reading prose passages. In two experiments, the open-book test yielded greater retrieval practice performance ($M = 0.81$) than the closed-book test ($M = 0.70$). When participants took a closed-book final test 1 week later, those who completed the open-book initial test still performed better than their closed-book counterparts ($M_{\text{open}} = 0.66$, $M_{\text{closed}} = 0.57$), although it is not clear whether those in the open-book initial test condition benefitted from additional reexposure to the original study material. Regardless, the findings from Cranney et al. (2009) and Agarwal et al. (2008) highlighted that a retrieval-based manipulation of retrieval practice performance (i.e., collaborative- vs. individual recall, open- vs. closed book) affected the size of the testing effect.

Other studies indicate that an encoding-based manipulation of retrieval practice performance can also influence the size of the testing effect. For example, de Lima et al. (2020) had participants study 20 easy and 20 difficult Swahili–Portuguese word pairs, and participants recalled more of the easy ($M = 0.58$) than the difficult pairs ($M = 0.35$) during retrieval practice. At the final test, relative to a restudy condition, the easy pairs produced a significantly larger testing effect ($M = 0.26$) than the difficult pairs ($M = 0.19$).

In a study that directly manipulated retrieval practice performance using an encoding-based manipulation, Racsmány et al. (2020) had participants study word pairs once, 3 times, or 6 times in three separate experiments, and then either practiced retrieval, restudied, or were not tested on the word pairs before a final test. As anticipated, more study repetitions led to greater retrieval practice performance on the initial test ($M_{1\times} = 0.25$, $M_{3\times} = 0.66$, $M_{6\times} = 0.76$). More importantly, relative to the no-test and restudy conditions, more study repetitions (3× or more) yielded a larger testing effect ($d_{1\times} = 0.96$, $d_{3\times} = 1.74$, $d_{6\times} = 1.61$),[2] thus underscoring a positive relationship between retrieval practice performance and the testing effect.

Although the above data appear to support a positive relationship, they should be interpreted with caution for several reasons. First, meta-analytic evidence is inherently observational/correlational, so the positive association between retrieval practice performance and the testing effect found in Rowland (2014) and Pan and Rickard (2018) could be explained by other factors (e.g., students from elite universities, who might be especially performance-oriented, might demonstrate greater retrieval practice performance and testing effects than students from other universities). A parallel concern also applies to Racsmány et al. (2020) because they had participants study items once in Experiment 1, 3 times in Experiment 2, and 6 times in Experiment 3, so the positive relationship here was established via a cross-experimental comparison. Second, the Cranney et al. (2009) study had a very small sample size (72 participants across four between-subjects conditions), and other factors regarding collaborative testing or open-book tests (Agarwal et al., 2008) may have contributed to differences in test performance. Lastly, although the easy items produced a larger testing effect than the difficult items in de Lima et al. (2020), the difference was very small ($d = 0.28$). We now consider evidence that greater retrieval practice performance does not necessarily lead to a larger testing effect.

Evidence for No Relationship or a Negative Relationship

Carpenter (2009) had participants study word pairs for which the cue words were either strongly or weakly related to the targets. As expected, during the initial test, participants recalled significantly more targets with strong cues than with weak cues ($M_{\text{strong}} = 0.96$, $M_{\text{weak}} = 0.91$). However, after a 5-min retention interval, the pattern was reversed, such that the targets previously tested with strong cues exhibited a smaller testing effect ($M = 0.12$) than those tested with weak cues ($M = 0.21$). In a subsequent experiment, Carpenter increased the number of study pairs to prevent retrieval practice performance from approaching ceiling, and participants still recalled more strong-cue targets ($M = 0.88$) than weak-cue targets ($M = 0.75$). Nevertheless, the two types of pairs yielded a comparable testing effect ($M_{\text{strong}} = 0.14$, $M_{\text{weak}} = 0.15$).

Fiechter and Benjamin (2018) had participants study Swahili–English word pairs either once or 3 times. During the practice phase, participants either completed a standard cued recall test, a diminishing-cue recall test,[3] or restudied the pairs. As expected, the thrice-studied pairs ($M_{\text{recall}} = 0.68$, $M_{\text{dim-recall}} = 0.87$) were recalled more often than the once-studied pairs ($M_{\text{recall}} = 0.38$, $M_{\text{dim-recall}} = 0.73$). However, when participants completed the final test 24 hr later, they produced a similar testing effect relative to restudy regardless of whether the pairs were studied thrice ($d_{\text{recall}} = 0.13$, $d_{\text{dim-recall}} = 0.37$) or once ($d_{\text{recall}} = -0.07$, $d_{\text{dim-recall}} = 0.28$).

In another study (Putnam & Roediger, 2013), after being presented with weakly related paired associates, participants either recalled the target words by typing them or by recalling them aloud. Participants recalled more items via typing ($M = 0.70$) than recalling aloud ($M = 0.60$). However, when they took a final test 2 days later,

[2] When reporting the size of the testing effect from existing studies, we report Cohen's d when possible. When the original article did not provide standard deviation or standard error, we report the mean difference in final test performance between the retrieval practice and restudy conditions.

[3] Across two rounds of initial test, participants were provided with fewer cue letters in the second round than the first.

the greater retrieval practice performance of the typing condition did not produce a larger testing effect than the aloud condition. This pattern held regardless of whether the control pairs were restudied ($d_{typing} = 0.96$, $d_{aloud} = 1.04$) or not ($d_{typing} = 1.45$, $d_{aloud} = 1.55$).

Kliegl et al. (2019) manipulated retrieval practice performance by varying the difficulty of the initial test. After studying weakly related word pairs (e.g., disappear—FADE), participants took an easy initial test (disappear—FA__) for some pairs and a difficult initial test for other pairs (disappear—F__). Participants recalled more targets in the easy test ($M = 0.73$) than in the difficult test ($M = 0.62$). However, upon returning 1 week later, they recalled more of the items that were queried on the difficult-initial test ($M = 0.29$) than those queried on the easy initial test ($M = 0.22$).

Smith and Karpicke (2014) had participants complete either a multiple-choice (MC) or short-answer (SA) initial test for prose materials. Across four experiments, MC consistently produced vastly superior retrieval practice performance relative to SA (average $d = 1.78$). However, on a 1-week delayed final recall test, compared to a no-test control condition, initial MC produced a similar testing effect as the initial SA in Experiment 1 ($d_{MC} = 1.52$, $d_{SA} = 1.34$) and a smaller testing effect than the initial short answer in Experiment 2 ($d_{MC} = 1.28$, $d_{SA} = 1.73$) and Experiment 4 ($d_{MC} = 0.89$, $d_{SA} = 1.11$). Experiment 3 did not include a no-test control condition, but it exhibited the same pattern of results as in Experiments 2 and 4 ($M_{MC} = .56$, $M_{SA} = .64$). Similar results have been observed in other studies that manipulated initial test format (Carpenter & DeLosh, 2006; Duchastel, 1981; Glover, 1989), although the pattern is not always consistent and might depend on whether or not feedback is given during the initial test (Kang et al., 2007; for reviews, see Adesope et al., 2017; Rowland, 2014; Yang et al., 2021).

In a study designed to explore the relationship between retrieval practice performance and the testing effect, Kanayama and Kasahara (2018) had participants restudy English–Japanese pairs that they had learned a week earlier—either just before or after the initial test. Not surprisingly, participants who restudied the pairs before the initial test ($M_{before} = 0.55$) produced markedly better retrieval practice performance than those who restudied after ($M_{after} = 0.02$). However, the opposite pattern was observed when participants were tested again on a final test. Here, the restudy-after participants outperformed the restudy-before participants, and this was true regardless of whether the final test was issued an hour ($M_{after} = 0.84$, $M_{before} = 0.53$) and a week later ($M_{after} = 0.55$, $M_{before} = 0.33$), despite that all participants had studied the pairs twice and taken a test on them once by the final test.

In a recent study, Gupta et al. (2022) examined whether differences in encoding-based episodic knowledge affected the size of the testing effect. Across three experiments, participants studied weakly associated pairs 1 or 4 times (Experiments 1 and 3), and 4 or 8 times (Experiment 2). Afterward, participants either received a cued recall test with feedback or restudied the pairs, and their memory was tested in a final test 48 hr later. In all three experiments, more study repetitions produced greater retrieval practice performance, but it did not produce a greater testing effect. To foreshadow, the design of these experiments are quite similar to our Experiments 2 and 3, although we also manipulated retention interval and feedback.

Lastly, we previously discussed a study by Cranney et al. (2009) that showed a positive relationship between retrieval practice performance and the testing effect. However, a different pattern of results arose in a similar study by Vojdanoska et al. (2009). Here, first-year psychology students completed an initial test for a PowerPoint presentation about adult development either individually or collaboratively. Like Cranney et al., collaborative recall led to vastly superior retrieval practice performance ($M = 0.67$) relative to individual recall ($M = 0.39$, $d = 1.92$). However, on a 1-week delayed final test, the two groups of students produced comparable testing effects ($M_{collaborative} = 0.10$, $M_{individual} = 0.13$). When considering the disparate results of collaborative versus individual testing on the testing effect across the two studies, Vojdanoska et al. suggested that the very high collaborative retrieval practice performance in Cranney et al. ($M = 0.94$) relative to Vojdanoksa et al. ($M = 0.67$) might explain why the former observed a collaborative advantage (but the latter did not). In addition to highlighting the disparate pattern of results in the literature, this notion illustrates the belief that greater retrieval practice performance should produce a larger testing effect.

Together, these results show that an increase in retrieval practice performance does not necessarily result in a larger testing effect; in fact, the opposite might occur. However, like the evidence for a positive relationship reviewed above, these data must be interpreted with caution. Because most existing studies were not explicitly designed to examine the relationship between retrieval practice performance and the testing effect, variations in retrieval practice performance coincided with changes in task demands. These changes might alter a learner's behavior or strategies (Ahn & Chan, 2022, 2023; Cho & Neely, 2017), which could have downstream consequences on final test performance and the testing effect that are independent of retrieval practice performance per se. For example, recalling materials collaboratively (Vojdanoska et al., 2009) can alter how students approach a retrieval task and promote processes that are otherwise absent during individual recall (e.g., collaborative inhibition; Blumen & Rajaram, 2008). Having participants restudy the material just before as opposed to after the initial test (Kanayama & Kasahara, 2018) can affect both retrieval practice performance and how learners reengage with the materials, because retrieval attempts can potentiate subsequent learning (Arnold & McDermott, 2013a; Chan et al., 2020; Chan, Manley, et al., 2018; Chan, Meissner, & Davis, 2018; Izawa, 1971; St. Hilaire et al., 2023).Moreover, changing the initial test format or difficulty (Kliegl et al., 2019; Smith & Karpicke, 2014) can affect the processes evoked during retrieval practice (Balota & Neely, 1980; Chan & McDermott, 2007; Cho & Neely, 2017; Shimizu & Jacoby, 2005; Whitten, 1978). Finally, the effect of increasing or reducing retrieval practice performance via an encoding manipulation (e.g., number of study trials, easy vs. difficult pairs) has not always produced consistent results, with some studies producing a positive relationship (de Lima et al., 2020; Racsmány et al., 2020) and others producing no relationship or a negative relationship (Carpenter, 2009; Fiechter & Benjamin, 2018). Thus, it is currently unclear from the extant data whether initial test performance influences the magnitude of the testing effect.

The Current Experiments

In this study, we aim to determine whether superior retrieval practice performance would translate to a more pronounced testing effect. To thoroughly assess our research question, we employed both a retrieval-based manipulation (in Experiments 1a and 1b) and an encoding-based manipulation (in Experiments 2a, 2b, 3a, and 3b) across six experiments. Experiments 1a and 1b varied retrieval practice performance with a response deadline manipulation. These

experiments were designed to hold encoding constant and varied retrieval practice performance via changes to the initial test environment. An additional contribution of Experiment 1 is that, despite the considerable body of literature on the testing effect, few studies have manipulated the duration of retrieval practice (cf., Chan, 2007; for semantic memory of trivia facts and remote associates, see Vaughn et al., 2017). Given that many students prepare for their exams under tight time constraints (Indig, 2005), they are likely to practice retrieval in a hurried manner, so it is important to understand if curtailing retrieval prematurely would diminish its advantages. In Experiments 2 and 3, we varied retrieval practice performance with a study trial manipulation—once versus twice in Experiment 2, and once versus 4 times in Experiment 3.

In addition to retrieval practice performance, we also manipulated whether participants received corrective feedback during the initial test. Prior studies have sometimes shown that the provision of feedback can eliminate the difference in retrieval practice performance (between a high-performing and a low-performing condition) on the final test (Kang et al., 2007), because feedback can preferentially improve performance in the low-performing condition. Given that students would likely learn the correct answers following retrieval practice in real learning scenarios, we implemented this design element to broaden the generalizability of our findings. Finally, in every pair of experiments, we also administered the final test following either a short (25 min) or a long (1 week) retention interval. This manipulation is important because the testing effect is sometimes greater following a substantial retention interval than after a brief one (Adesope et al., 2017; Rowland, 2014, but for a thoughtful consideration against overinterpreting this effect, see Karpicke et al., 2014), so this manipulation further enhances the generalizability of our findings.

Transparency and Openness

These experiments were not preregistered, and sample sizes were determined based on convention at the time of data collection (2014–2017). Therefore, instead of a power analysis, we report a sensitivity analysis for each experiment. All data exclusions, manipulations, and measures in the experiments fit the Journal Articles Reporting Standards (Appelbaum et al., 2018). We analyzed our data with JASP (JASP Team, 2023), and all data and materials are available on the Open Science Framework (OSF) (Yurtsever et al., 2024) at https://osf.io/st596. All experiments reported here were approved by the Institutional Review Board at Iowa State University.

Experiment 1a

Method

Design and Participants

In Experiment 1a, initial test condition (no-test, short-deadline, medium-deadline, or long-deadline) was manipulated within subjects, and feedback was manipulated between subjects. Participants in the feedback condition only received feedback for items that appeared on the initial test. As such, it was not possible to employ a full factorial design.

Fifty-seven participants from Iowa State University participated in the experiment for partial course credit. Data from nine participants were excluded from analyses as they indicated that English was not their native language, resulting in 24 participants in each between-subjects condition (feedback or no-feedback). A sensitivity analysis indicated that, at .80 power, a sample size of 24 permitted detection of a within-subjects effect size of $f = 0.27$ ($d = 0.53$),[4] and a sample size of 48 (when collapsed across the feedback variable) permitted detection of a within-subjects effect size of $f = 0.19$ ($d = 0.37$).

Materials and Procedure

Participants studied four lists of category-exemplar pairs (e.g., weapon: bomb; animal: bear). Each list had 24 pairs, which consisted of six exemplars from four categories (see the OSF page for the full set of materials). The taxonomic frequency of exemplars across the 16 categories did not differ significantly ($M = 0.14$, $SD = 0.03$), $F(15, 80) = 0.29$, $p = .995$. Participants studied each word pair for 4 s with a 300-ms interstimulus interval (ITI). After studying a list, participants would complete either math problems (i.e., no test) or an initial test with a short-, medium-, or long-response deadline. They would then study the next list and complete a different task following study.[5]

During each trial with an initial recall test, participants saw the category name and the first two letters of the exemplar (e.g., animal: do__). They were instructed to type just the part that fit the blank or the entire word. The recall trial ended when the appropriate response deadline had elapsed—3 s for the short response deadline, 5.5 s for a medium deadline, and 8 s for a long deadline. Participants were informed about the response deadline of the upcoming trials at the start of each initial test. Participants in the feedback condition were then presented with the complete pair for 2 s, whereas participants in the no-feedback condition proceeded immediately to the next initial test trial. For the no-test condition, participants completed math problems instead of an initial test. The duration of the math task (72 s) matched the short deadline recall task. The math task comprised six algebra problems, and participants had 12 s to answer each question (e.g., [17–12 + 3]/8).

The study and initial test phase spanned roughly 20 min. Participants then had a 25-min retention interval, during which they completed the automatic reading span task (RSPAN; Unsworth et al., 2005). In this task, participants determined whether a series of sentences made sense (e.g., "After yelling at the game, I knew I would have a tall voice"). A letter appeared after each sentence. After a varying number of sentences, participants were prompted to recall the letters in the presented order. If participants completed the entire RSPAN task before the 25-min interval expired, they played the video game Tetris for the residual time. The RSPAN was terminated if participants did not finish within 25 min.

After the 25-min retention interval, participants completed a final recall test on all 96 studied pairs. This final test was self-paced. Participants were explicitly told the following: "Unlike the

[4] For the sensitivity analysis, we used G*Power ANOVA: repeated measures, within factors function. Because we were interested in the main effect of response deadline on (a) retrieval practice performance and (b) testing effect, we conducted our sensitivity analysis for this test, which contained a single within-subjects variable with three levels (short, medium, and long). We did not include the between-subjects factor feedback in the sensitivity analysis because we did not predict an interaction between response deadline and feedback.

[5] Given the multi-list learning task structure, other indirect benefits of testing (e.g., test-potentiated new learning; Chan, Manley, et al., 2018; Chan, Meissner, & Davis, 2018) might have impacted learning of subsequent lists. However, list order analyses for our data in Experiments 1a and 1b indicated no evidence for such order effects (see OSF for list order data).

previous test, there is no time limit on this one. Please answer as many questions as you can, and take as long as you need on each question." No feedback was given during this final test phase. Finally, participants completed a brief demographic survey and were debriefed.

Results and Discussion

We conducted our conventional analyses using two-tailed tests with an alpha level of .05. Effect sizes were indexed by Cohen's *d* and partial eta squared. We also provide Bayes factors (BF) for each analysis. In addition to providing support for the alternative hypothesis over the null, BF can quantify the level of evidence in support of the null hypothesis, which allows us to evaluate the hypothesis that the response deadline would not affect the magnitude of the testing effect. All Bayesian analyses were performed in JASP using the default priors. Specifically, a point-zero prior was used for H_0, and a Cauchy distribution with a zero center and $r = \pm .707$ for its interquartile range was used for H_1 (van Ravenzwaaij & Wagenmakers, 2022). We report BF_{10} when the data were more probable under H_1 relative to H_0 and BF_{01} when the data were more probable under H_0 relative to H_1. Therefore, a larger BF always indicates more support for the hypothesis (whether H_0 or H_1) being discussed. For example, if a set of data is more probable under H_1 than H_0, we report BF_{10}; alternatively, if a set of data offers stronger support for H_0 than H_1, we report BF_{01}. We adopted this approach to simplify readers' understanding (e.g., a $BF_{10} = 0.125$ is likely more difficult to interpret than $BF_{01} = 8$). Another important point to note is that it is much easier to obtain very large BFs (e.g., in the thousands or millions) supporting H_1 than similar magnitudes of BFs supporting H_0. One reason for this is because the default prior for H_1 includes zero (and therefore H_0). As Brysbaert (2019) highlighted, a BF_{01} of 10 is unreachable for most research (at least with the default priors), so it is important to keep this disparity in mind when interpreting the BF_{01} and BF_{10} reported here.

Retrieval Practice Performance From the Initial Test

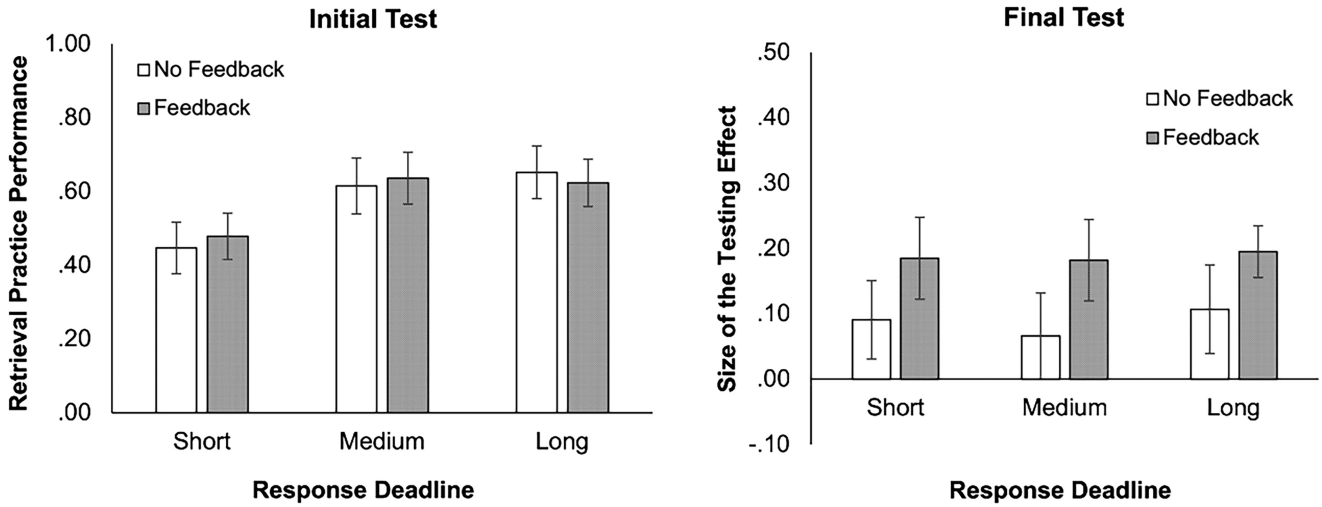
A one-way repeated-measures analysis of variance (ANOVA) revealed a very large effect of response deadline on retrieval practice performance, $F(2, 94) = 28.14$, $p < .001$, $\eta^2 = .37$, $BF_{10} = 4.229 \times 10^7$, as evidenced by the eta-squared statistics and Figure 1. Moreover, the data strongly supported H_1 over the null. Post hoc comparisons using Holm correction demonstrated that performance was significantly worse under the short deadline ($M = 0.46$, $SD = 0.17$) compared to both the medium deadline ($M = 0.63$, $SD = 0.18$), $t(47) = -6.25$, $p < .001$, $BF_{10} = 20{,}587.23$, and the long deadline ($M = 0.64$, $SD = 0.17$), $t(47) = -6.72$, $p < .001$, $BF_{10} = 958{,}462.55$. However, there was no significant difference in performance between the medium and long deadlines, with the BF indicating that the null was 5 times more likely than the alternative hypothesis, $t(47) = -0.46$, $p = .643$, $BF_{01} = 5.64$. Therefore, we successfully manipulated retrieval practice performance via the response deadlines, although performance plateaued at the medium deadline.

Final Test Performance and the Testing Effect

Given the success of our deadline manipulation, if retrieval practice performance is positively related to the magnitude of the testing effect, then we should observe a larger testing effect for items practiced under the medium and long deadlines compared to those practiced under the short deadline. To calculate the testing effect, we subtracted proportion recalled of items not initially tested from the proportion recalled of items tested under short, medium, and long deadlines for each participant. For instance, during the final test, if a participant recalled .85 of the targets initially tested with a long deadline, .70 of the targets with a medium deadline, .50 of the targets with a short deadline, and .40 of the targets omitted from the initial test, then this participant would have a testing effect of .45 for long deadline items, .30 for medium deadline items, and .10 for short deadline items. Subtracting away the no-test items' recall probability allowed us to analyze our final test data using a factorial design.

Figure 1

Retrieval Practice Performance and the Testing Effect as a Function of Response Deadline and Feedback in Experiment 1a



Note. The left and right panels have different scales for the *y* axes. Because the testing effect is a difference score between the tested and control items, it can be negative if participants recall more control items than tested items. Error bars are descriptive 0.95 confidence intervals.

One might be concerned about the reliability of difference scores, but because the scores across conditions were subtracted from the same baseline (i.e., there was only one set of no-test items), the subtraction is a linear transformation and would not affect the outcomes of the inferential analyses (relative to using raw proportions of recall). Furthermore, the notion of difference scores being less reliable than raw scores is not justified (Nickerson & Brown, 2019; Overall & Woodward, 1975; Thomas & Zumbo, 2011; Trafimow, 2015). However, interested readers will find the raw recall probabilities of the final test in Table 1.

We analyzed the magnitude of the testing effect observed on the final test with a 3 (response deadline: short, medium, long) × 2 (feedback, no-feedback) ANOVA. Most notably, response deadline did not influence the testing effect, $F(2, 92) = 0.87$, $p = .421$, $\eta_p^2 = .02$, $BF_{01} = 7.10$. Specifically, the mean magnitude of the testing effect at the short, medium, and long deadlines was .14, .12, and .15, respectively. Unlike the powerful effect observed during the initial test, the BF shows substantial support for the null over the alternative hypothesis. The main effect of feedback, however, was significant, $F(1, 46) = 7.28$, $p = .010$, $\eta_p^2 = .14$, $BF_{10} = 5.16$, which shows that the provision of feedback boosted the magnitude of the testing effect ($M = 0.19$ with feedback, $M = 0.09$ without feedback). However, we caution against overinterpreting this main effect. Note that the testing effect for no-feedback items was based on a comparison between nontested, no-feedback items and tested, no-feedback items. However, the testing effect for feedback items was based on a comparison between nontested, no-feedback items and tested-feedback items (nontested items cannot receive feedback). Therefore, the feedback effect here stemmed from a combination of both the reexposure from feedback and the beneficial effects of posttest processing of the feedback (Arnold & McDermott, 2013b; Mulligan et al., 2022).

The interaction between these two factors was also not significant, $F(2, 92) = 0.27$, $p = .761$, $\eta_p^2 = .01$, $BF_{01} = 6.735$. As Figure 1 shows, the provision of feedback did not influence the (null) effect of response deadlines on the magnitude of the testing effect (the testing effect was $M_{short} = 0.19$, $M_{medium} = 0.18$, and $M_{long} = 0.19$) or not ($M_{short} = 0.09$, $M_{medium} = 0.07$, and $M_{long} = 0.11$). Therefore, despite the considerable influence of the response deadlines on retrieval practice performance, these performance differences did not persist to the final test. This conclusion is further buttressed by the Bayesian factors. Indeed, even if we split the data in half and examined the effects of response deadline on the testing effect separately for the no-feedback and feedback groups, the BF still favored the null over the alternative hypothesis, BF_{01} for no-feedback = 4.50, BF_{01} for feedback = 7.57.

Overall, the data from Experiment 1a suggest that an experimental manipulation of retrieval practice performance via response deadline does not affect the size of the testing effect. However, given the novelty of this finding, we sought to replicate and extend it in Experiment 1b to a final test administered after a retention interval of 1 week.

Experiment 1b

Method

Participants

Sixty-eight students from Iowa State University participated for partial course credit. The data from five participants were excluded from analyses because they indicated that English was not their native language. Data analysis was thus conducted on 31 participants in the feedback condition and 32 in the no-feedback condition. A sensitivity analysis indicates that, at .80 power, a sample size of 31 permitted detection of a within-subjects effect size of $f = 0.23$ ($d = 0.47$), and a sample size of 63 (when collapsed across the feedback variable) permitted detection of a within-subjects effect size of $f = 0.16$ ($d = 0.32$).

Materials and Procedure

The design and procedure of Experiment 1b mirrored Experiment 1a with two modifications. First, we shortened the medium and long deadlines from 5.5 and 8 s in Experiment 1a to 4 and 7 s. This change was made because retrieval practice performance plateaued at 5.5 s in Experiment 1a. The short deadline

Table 1
Proportion Correct on the Final Test Across Experiments

	No feedback		Feedback			No feedback		Feedback	
Condition	*M*	*SD*	*M*	*SD*	Condition	*M*	*SD*	*M*	*SD*
Experiment 1a					Experiment 1b				
No initial test	0.55	0.15	0.61	0.13	No initial test	0.39	0.19	0.40	0.17
Short deadline	0.64	0.17	0.79	0.14	Short deadline	0.50	0.23	0.62	0.19
Medium deadline	0.62	0.20	0.79	0.15	Medium deadline	0.52	0.23	0.56	0.17
Long deadline	0.66	0.20	0.80	0.12	Long deadline	0.50	0.24	0.61	0.17
Experiment 2a					Experiment 2b				
No initial test-1×	0.66	0.14			No initial test-1×	0.53	0.14		
No initial test-2×	0.74	0.13			No initial test-2×	0.59	0.15		
Tested-1×	0.78	0.14	0.89	0.12	Tested-1×	0.73	0.14	0.75	0.20
Tested-2×	0.83	0.16	0.94	0.11	Tested-2×	0.70	0.19	0.75	0.19
Experiment 3a					Experiment 3b				
No initial test-1×	0.32	0.16			No initial test-1×	0.07	0.06		
No initial test-4×	0.55	0.21			No initial test-4×	0.11	0.08		
Tested-1×	0.46	0.22	0.73	0.20	Tested-1×	0.18	0.14	0.18	0.15
Tested-4×	0.70	0.20	0.78	0.17	Tested-4×	0.24	0.17	0.26	0.13

remained identical to Experiment 1a at 3 s. Second, rather than completing the entire experiment in a single session, participants were dismissed after the initial test phase. A week later, they returned to the lab, did RSPAN and Tetris for 25 min, and then took the final test.

Results and Discussion

Retrieval Practice Performance

Similar to Experiment 1a, response deadline had a powerful effect on retrieval practice performance, $F(2, 124) = 25.03$, $p < .001$, $\eta^2 = .29$, $BF_{10} = 1.280 \times 10^7$. Post hoc comparisons using Holm correction showed that participants recalled fewer items under the short deadline ($M = 0.50$, $SD = 0.20$) than the medium deadline ($M = 0.57$, $SD = 0.18$), $t(62) = -3.47$, $p < .001$, and they recalled fewer items under the medium deadline than the long deadline ($M = 0.64$, $SD = 0.19$), $t(62) = -3.60$, $p < .001$. Unlike Experiment 1a, in which retrieval practice performance peaked at the medium deadline, the revised response deadlines successfully differentiated retrieval practice performance at each level.

Final Test Performance

Contrary to results from the initial test, response deadline did not yield a significant main effect on the testing effect, $F(2, 122) = 0.29$, $p = .752$, $\eta_p^2 = .01$, $BF_{01} = 15.02$. Indeed, the BF shows strong support for the null. The main effect of feedback was significant, $F(1, 61) = 6.70$, $p = .012$, $\eta_p^2 = .10$, $BF_{10} = 3.80$, and the interaction was not, $F(2, 122) = 1.50$, $p = .227$, $\eta_p^2 = .02$, $BF_{01} = 3.09$ (see Figure 2). Once again, a longer response deadline did not translate to an increase in the testing effect regardless of whether participants were provided with feedback during the initial test ($M_{short} = 0.22$, $M_{medium} = 0.16$, $M_{long} = 0.21$, $BF_{01} = 3.12$) or not ($M_{short} = 0.11$, $M_{medium} = 0.13$, $M_{long} = 0.11$, $BF_{01} = 8.83$).

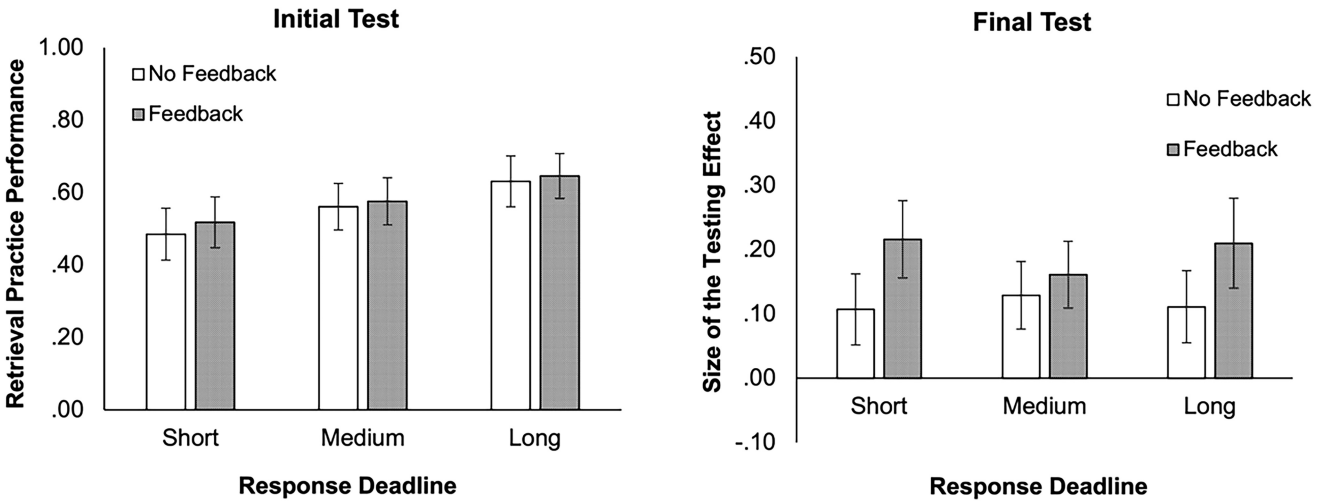
Summary of Experiment 1

Experiment 1 showed that, despite the potent effect of response deadline on retrieval practice performance, it had little influence on the magnitude of the testing effect at either a 25-min or 1-week retention interval. Additionally, this pattern persisted whether participants received feedback during the initial test or not. These findings suggest that our retrieval-based manipulation of retrieval practice performance did not affect the testing effect. Moreover, they showed that learners under time constraints might still reap the full benefit of retrieval practice, even if they do not have enough time to recall all of the retrievable items. However, one might wonder if our response deadline manipulation failed to affect the testing effect because it only affected the output of retrieval, rather than the product of retrieval per se. Specifically, the short response deadline might actually be long enough for participants to mentally retrieve the same number of items as the long response deadline, and the reduction in retrieval practice performance simply reflects insufficient time for participants to physically type all of the retrieved items. In other words, the effect of response deadline on retrieval practice performance was a methodological artifact rather than a genuine memory phenomenon.

Although this alternative explanation seems plausible, it is difficult to falsify. We are not aware of any way to verify what is or is not retrieved inside participants' minds (outside of what they can output), so the same argument could be leveled against any method that requires external responses from participants (e.g., participants might speak faster than they can type, but one can still suggest that participants might not have enough time to speak), which renders the argument unfalsifiable. However, regardless of whether this "artifact explanation" is substantive, we sought to further investigate the relationship between retrieval practice performance and the testing effect in another way. In Experiment 1, we manipulated retrieval practice performance using a retrieval-based manipulation. In Experiment 2, we employed an encoding-based manipulation such that half of the items were presented once and half were presented twice.

Figure 2

Retrieval Practice Performance and the Testing Effect as a Function of Response Deadline and Feedback in Experiment 1b



Note. The left and right panels have different scales for the *y* axes. Because the testing effect is a difference score between the tested and control items, it can be negative if participants recall more control items than tested items. Error bars are descriptive 0.95 confidence intervals.

Experiment 2a

Method

Design and Participants

Experiment 2a used a 2 (encoding presentation: 1×, 2×) × 2 (retrieval practice condition: test, no test) design, with the additional variable of feedback nested inside the tested items, such that half of the tested items received feedback and half did not. All variables, including feedback, were manipulated within subjects.

Seventy-one participants at Iowa State University participated in exchange for partial course credit. The data from 16 participants were omitted from analyses because they indicated that English was not their primary language. Data from an additional 14 participants were removed due to an experimenter error. Thus, the final analysis included data from 41 participants. A sensitivity analysis showed that this sample size permitted the detection of a within-subjects effect size of $f = 0.22$ ($d = 0.44$).[6]

Materials and Procedure

Because we manipulated all variables within subjects, we required a larger set of items for Experiments 2 and 3 relative to Experiment 1. To this end, we constructed four lists of 24 weakly associated cue-target pairs (e.g., Chisel-Sculpture; see the OSF page for the full set of materials). Forty-nine items were taken from Kornell et al.'s (2009) weakly associated stimuli, $M_{\text{forward-associative-strength (FAS)}} = 0.01$, $M_{\text{backward-associative-strength (BAS)}} = 0.05$. We created 47 additional cue-target pairs and matched these to the Kornell et al. items ($M_{\text{FAS}} = 0.01$, $M_{\text{BAS}} = 0.05$). BAS was held constant at .05 across all pairs, but FAS varied across pairs. Items were sorted randomly into four lists, and the presentation order of each list was randomized. Within each list, half of the items appeared once and half appeared twice, with the assignment of item to presentation condition counterbalanced across participants.

At the beginning of the study phase, participants were instructed to memorize as many word pairs as possible and that they would receive the left side of the pair as a cue to recall the target. Each cue-target pair was presented on the screen for 2.5 s, separated by a 300-ms ITI. Each list was followed immediately by a test. Half of the items from each list were tested, with half of these tested items receiving feedback. These conditions were counterbalanced across subjects, ensuring that an item was either tested or not tested, and if tested, it either received feedback or did not. On a given test trial, the cue was presented with a two-letter target stem (e.g., Chisel-Sc_____) for 6 s. For items on which feedback was provided, the correct answer appeared for two seconds following the completion of the test trial. Note that all item types were randomly intermixed within each study and test list. That is, within each study list, some pairs were studied once and others were studied twice, and the presentation order of all pairs was randomized within the study list. Likewise, in each initial test phase, participants would be asked to recall six studied-once pairs and six studied-twice pairs, with half of them being followed by the presentation of immediate feedback. The presentation order of these test trials within each initial test phase was also randomized.

After all four lists had been tested, participants completed the RSPAN task for 25 min. Immediately following this task, they proceeded to the self-paced final test. All 96 items were tested in a random order. We made the final test more difficult relative to the initial test by presenting participants with the cue and only a one-letter target stem (e.g., "Chisel-S_____").

Results and Discussion

Retrieval Practice Performance

The initial test data mirrored those from Experiment 1 (see Figure 3). Most importantly, encoding presentations had a large effect, $t(40) = 6.95$, $p < .001$, $d = 1.09$, $\text{BF}_{10} = 6.09 \times 10^5$. Encoding twice ($M = 0.88$) was associated with significantly better retrieval performance than encoding once ($M = 0.76$).

Final Test Performance and the Testing Effect

We now report the results of a 2 (encoding presentation: 1× or 2×) × 2 (feedback or no-feedback) repeated-measures ANOVA, with the size of the testing effect as the dependent measure. In Experiment 2, the size of the testing effect was calculated by subtracting final recall for nontested words from final recall for tested words in the respective conditions (e.g., 2× tested words − 2× nontested words). Contrary to the retrieval practice data, encoding presentation did not produce a significant main effect on the testing effect, $F(1, 40) = 1.26$, $p = .269$, $\eta_p^2 = .03$, $\text{BF}_{01} = 2.17$. One might notice that the BF provided very weak support for the null, but that is due to the small difference in the size of the testing effect in the negative direction ($M_{1\times} = 0.18$, $M_{2\times} = 0.14$, $d = -0.18$). That is, the condition that yielded greater retrieval practice performance produced a numerically (but not significantly) smaller testing effect. We do not interpret any reversals in the present study because (a) they are not significant and (b) we did not predict them. Moreover, the reversal here might be due in part to the near-ceiling recall performance for the items that received feedback (see Table 1). Because we are investigating a directional research question, an alternative Bayesian test is to set a directional prior for H_1 such that the study-twice condition is expected to produce a larger testing effect than the study-once condition. With this one-sided prior for H_1, our results provide strong evidence for the null, $\text{BF}_{0+} = 11.73$.

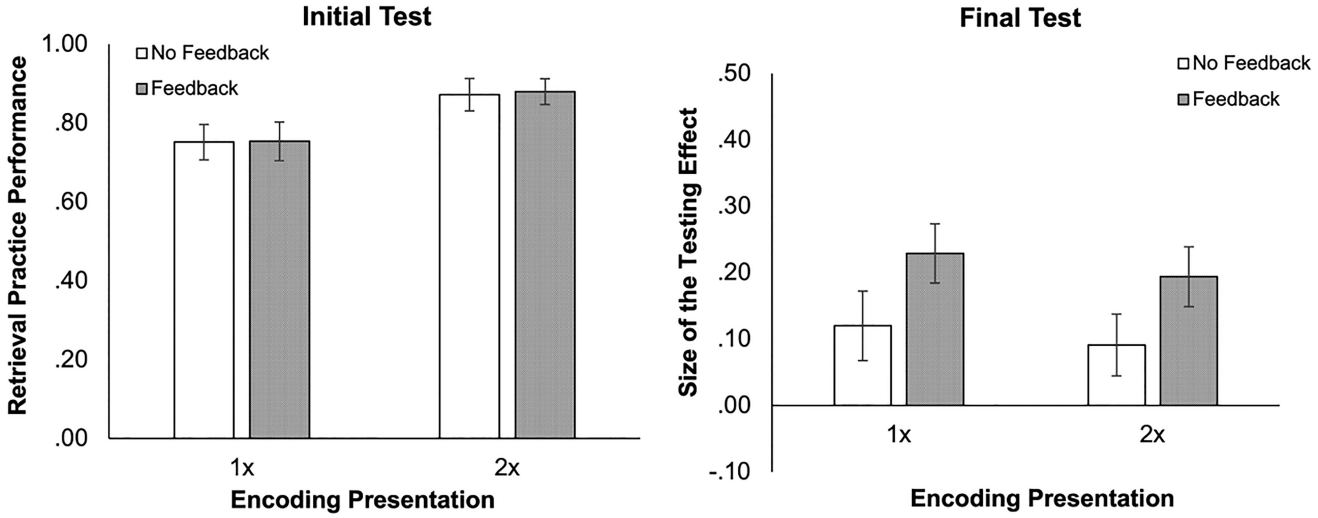
The main effect of feedback was significant, $F(1, 40) = 34.73$, $p < .001$, $\eta_p^2 = .47$, $\text{BF}_{10} = 1.23 \times 10^4$, and the interaction was not, $F(1, 40) = 0.04$, $p = .834$, $\eta_p^2 = .01$, $\text{BF}_{01} = 4.52$. When we examined the effect of encoding presentations on the testing effect separately for items practiced with ($M_{1\times} = 0.23$, $M_{2\times} = 0.19$, $\text{BF}_{01} = 3.04$) and without feedback ($M_{1\times} = 0.12$, $M_{2\times} = 0.09$, $\text{BF}_{01} = 4.37$), they again led to the same conclusion—an increase in retrieval practice performance did not lead to a larger testing effect.

In sum, the data from Experiment 2a mirrored those from Experiment 1a, despite the two experiments having different manipulations of retrieval practice. We now turn to Experiment 2b, in which we introduced a 1-week retention interval similar to Experiment 1b.

[6] Because we are not predicting any interactions, this sensitivity analysis was designed for the main effect of encoding presentation—a two-level within-subjects variable.

Figure 3

Retrieval Practice Performance and the Testing Effect as a Function of Encoding Presentation and Feedback in Experiment 2a



Note. The left and right panels have different scales for the *y* axes. Because the testing effect is a difference score between the tested and control items, it can be negative if participants recall more control items than tested items. Error bars are descriptive 0.95 confidence intervals.

Experiment 2b

Method

Participants

Forty-four participants at Iowa State University participated in exchange for partial course credit. Data from five were removed from analyses because they failed to return for the second session of the experiment, and four were removed because English was not their native language. Data analysis was conducted with 35 participants, which permitted the detection of a within-subjects effect size of $f = 0.24$ ($d = 0.49$).

Materials and Procedure

The design and procedure of Experiment 2b were the same as Experiment 2a, except that participants returned to the lab 1 week after they had completed the initial test phase. Upon returning to the lab, participants completed the RSPAN task and then the final test.

Results and Discussion

Retrieval Practice Performance

Similar to Experiment 2a, participants recalled more target words after having studied them twice ($M = 0.86$) than once ($M = 0.77$), $t(34) = 5.85$, $p < .001$, $d = 0.99$, $BF_{10} = 1.30 \times 10^4$, as shown in Figure 4.

Final Test Performance and the Testing Effect

Consistent with the main finding from Experiment 2a, encoding presentation did not produce a significant main effect on the testing effect, $F(1, 34) = 3.64$, $p = .065$, $\eta_p^2 = .10$, $BF_{01} = 1.16$. Indeed, if anything, the effect was again in the negative direction, such that items studied twice ($M = 0.13$) produced a smaller testing effect than items studied once ($M = 0.21$), $d = 0.32$, $BF_{0+} = 14.53$.[7] This reversal is more notable than the one in Experiment 2a because it was not obscured by ceiling effects in test accuracy (see Table 1). When we examined the pairwise comparisons, encoding presentations again did not increase the testing effect for both items that received feedback ($M_{1\times} = 0.21$, $M_{2\times} = 0.15$, $BF_{01} = 2.94$, $BF_{0+} = 11.06$) or not ($M_{1\times} = 0.20$, $M_{2\times} = 0.11$, $BF_{10} = 1.13$, $BF_{0+} = 15.28$).

Unlike the results of Experiment 2a, the main effect of feedback was not significant, $F(1, 34) = 3.13$, $p = .086$, $\eta_p^2 = .08$, $BF_{01} = 2.49$. The beneficial effects of the extra presentation from feedback had apparently dissipated across the 1-week retention interval ($M_{\text{no-feedback}} = .15$, $M_{\text{feedback}} = .18$). Lastly, the interaction between encoding presentation and feedback was also not significant, $F(1, 34) = 0.38$, $p = .540$, $\eta_p^2 = .01$, $BF_{01} = 3.29$.

Summary of Experiment 2

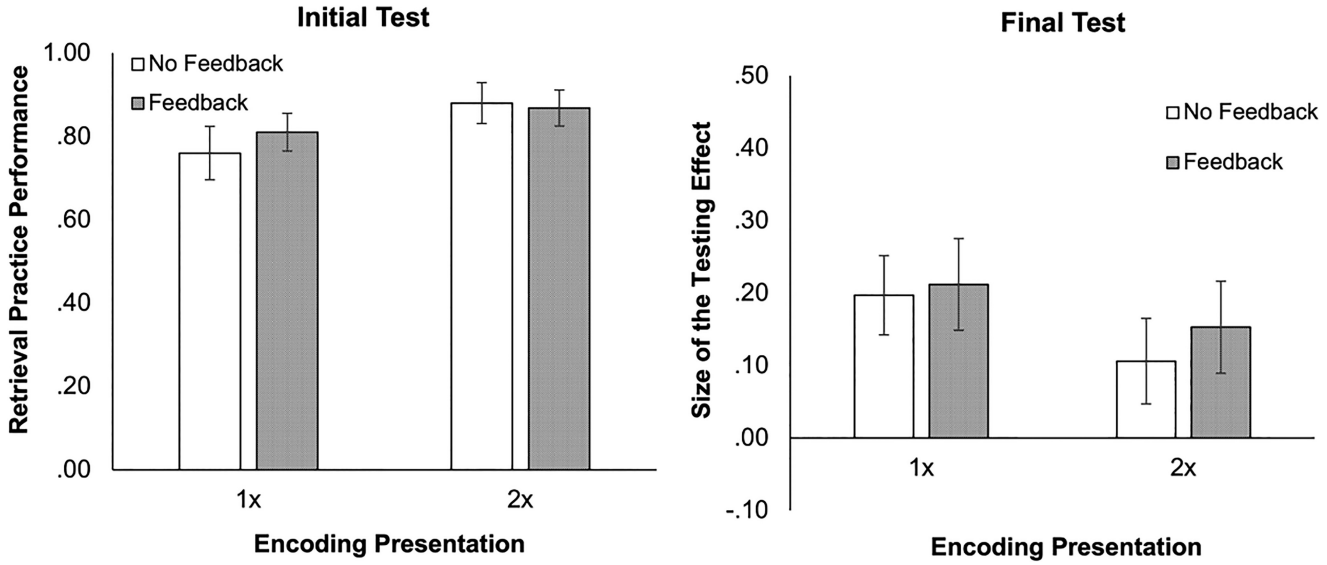
In Experiment 2, we successfully varied retrieval practice performance using an encoding-based manipulation. Despite the powerful effect that two encoding trials had on retrieval practice performance relative to a single encoding trial ($d = 1.09$ for Experiment 2a and $d = 0.99$ for Experiment 2b), it did not increase the size of the testing effect either at the 25 min ($d = -0.18$ for Experiment 2a) or 1-week retention interval ($d = -0.32$ for Experiment 2b). Indeed, if anything, the testing effect became slightly smaller in the encoding-twice condition.

Although these data are consistent with those from Experiment 1, one concern is that the effect of our encoding-based manipulation was perhaps still not powerful enough to elicit a difference on the final test. Specifically, although the effect of encoding presentation was substantial in Experiments 2a and 2b, it occurred in the context

[7] When a Bayesian analysis is one-sided, the notation for the Bayes Factor becomes BF_{0+} or BF_{+0}.

Figure 4

Retrieval Practice Performance and the Testing Effect as a Function of Encoding Presentation and Feedback in Experiment 2b



Note. The left and right panels have different scales for the *y* axes. Because the testing effect is a difference score between the tested and control items, it can be negative if participants recall more control items than tested items. Error bars are descriptive 0.95 confidence intervals.

of relatively high retrieval practice performance ($M_{1\times} = 0.76$, $M_{2\times} = 0.87$, across the two experiments). It is possible that the putative positive association between retrieval practice performance and the testing effect only occurs when the difference in retrieval practice performance is very large (e.g., although the standardized effect size was large, the raw difference was modest at about .11) and the baseline retrieval practice performance is considerably lower (e.g., $M_{1\times} = 0.76$ might be too high as a baseline performance). In Experiment 3, we sought to address these possibilities by (a) strengthening the magnitude of the encoding manipulation while (b) reducing baseline retrieval practice performance.

Experiment 3a

Method

Participants

Forty-four students at Iowa State University participated in exchange for partial course credit. Data from seven participants were excluded from the analysis due to an experimenter error, and two because English was not their native language. Data analysis was conducted with 35 participants. With .80 power, we detected an effect size of $f = 0.24$ ($d = 0.49$) with the given sample size.

Materials and Procedure

The design and procedure were the same as Experiment 2 but with three important differences. First, items were presented either once (1×) or 4 times (4×) during the initial encoding of the lists. Second, in an effort to reduce retrieval practice performance, items on the initial test were tested with a one-letter stem (e.g., Chisel-S_____) instead of a two-letter stem. Likewise, when items were tested on the final test, no stems were given (e.g., Chisel-_____).

Results and Discussion

Retrieval Practice Performance

As can be seen in Figure 5, the revised encoding manipulation had a massive effect on retrieval practice performance, $t(34) = 11.55$, $p < .001$, $d = 1.95$, $BF_{10} = 3.00 \times 10^{10}$. Items studied 4 times ($M = .78$) were recalled far more frequently than those studied once ($M = .53$), indicating that the increase in the number of presentations here did serve to produce a large performance difference between conditions than in Experiment 2.

Final Test Performance and the Testing Effect

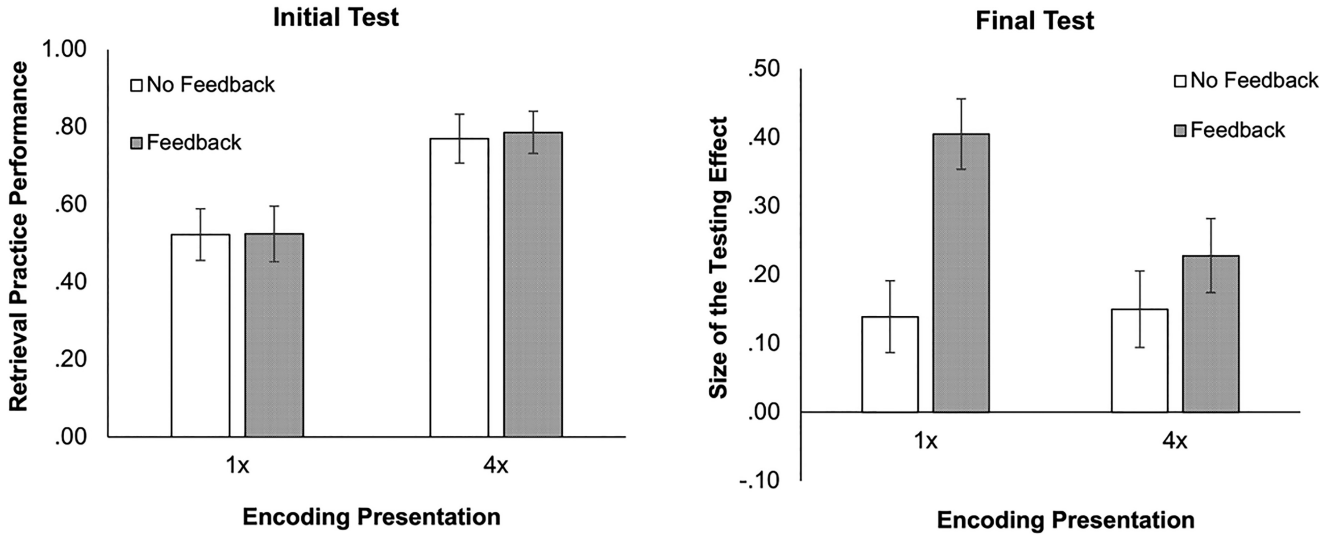
We now report results from the final test. Unlike each of the previous experiments, the number of encoding presentations had a significant main effect on the testing effect, $F(1, 34) = 7.24$, $p = .011$, $\eta_p^2 = .18$, $BF_{10} = 4.17$. However, it was in the negative direction, such that items studied 4 times actually produced a smaller testing effect ($M = 0.19$) than those studied once ($M = 0.28$). Note, however, that the BF only shows weak support for H_1 over H_0, so this reduction in the testing effect should be interpreted with caution.

Unsurprisingly, the main effect of feedback was significant, $F(1, 34) = 101.26$, $p < .001$, $\eta_p^2 = .75$, $BF_{10} = 9.55 \times 10^6$. The interaction was also significant, $F(1, 34) = 30.13$, $p < .001$, $\eta_p^2 = .47$, $BF_{10} = 1.06 \times 10^5$. A closer examination shows that number of study presentations had virtually no influence on the testing effect for items without feedback ($M_{1\times} = 0.14$, $M_{4\times} = 0.15$, $BF_{01} = 5.29$). However, for items with feedback, those studied 4 times actually demonstrated a smaller testing effect than those studied once ($M_{1\times} = 0.41$, $M_{4\times} = 0.23$, $BF_{10} = 1{,}194.61$), a negative association between retrieval practice performance and the testing effect.

Overall, the data in Experiment 3a are largely consistent with those from the previous experiments. We now report our final

Figure 5

Retrieval Practice Performance and the Testing Effect as a Function of Encoding Presentation and Feedback in Experiment 3a



Note. The left and right panels had different scales for the *y* axes. Because the testing effect is a difference score between the tested and control items, it can be negative if participants recall more control items than tested items. Error bars are descriptive 0.95 confidence intervals.

experiment in this article, in which we again varied retrieval practice performance using a 1× versus 4× manipulation, but the final test occurred following a 1-week retention interval.

Experiment 3b

Method

Participants

Fifty-three students at Iowa State University participated in exchange for partial course credit. Data from seven participants were excluded from analyses because English was not their native language, and nine because they did not complete the second session. Data analysis was thus conducted with 37 participants ($f = 0.24$, $d = .47$).

Materials and Procedure

The design and procedure were identical to Experiment 3a, except that participants returned for the RSPAN and the final test 1 week after the first session.

Results and Discussion

Retrieval Practice Performance

Figure 6 once again shows the potent effect of encoding presentation on retrieval practice performance, $t(36) = 14.22$, $p < .001$, $d = 2.34$, $BF_{10} = 2.44 \times 10^{13}$. Specifically, items studied 4 times ($M = 0.79$) were recalled far more often than those studied once ($M = 0.51$).

Final Test Performance and the Testing Effect

Overall, the results from this experiment were most similar to those from Experiment 2b (which had the same 1-week retention interval but a slightly less potent encoding manipulation). Specifically, encoding presentation did not have a significant main effect on the testing effect, $F(1, 36) = 1.62$, $p = .211$, $\eta_p^2 = .04$, $BF_{01} = 1.94$. The difference in recall between items encoded once and 4 times was negligible but in the positive direction ($M_{1\times} = 0.12$, $M_{4\times} = 0.15$), $d = 0.20$, and the pairwise comparisons led to the same conclusion for items tested without feedback ($M_{1\times} = 0.11$, $M_{4\times} = 0.13$, $BF_{01} = 3.83$) and for items tested with feedback ($M_{1\times} = 0.12$, $M_{4\times} = 0.15$, $BF_{01} = 2.89$). The main effect of feedback was also not significant, $F(1, 36) = 0.36$, $p = .554$, $\eta_p^2 = .01$, $BF_{01} = 3.87$. Indeed, like Experiment 2b and in contrast to Experiment 3a, the beneficial effects of feedback diminished over the week-long retention interval. Lastly, there was no interaction between encoding presentation and feedback, $F(1, 36) = 0.08$, $p = .776$, $\eta_p^2 < .01$, $BF_{01} = 4.27$.

Summary of Experiment 3

In Experiment 3, we aimed to implement a stronger encoding manipulation and to lower the baseline retrieval practice performance by presenting a one-letter stem as the cue, and we accomplished both of these goals. Specifically, more encoding opportunities resulted in a greater mean difference between the two encoding conditions on the initial test ($d_{E3} = 2.15$ vs. $d_{E2} = 1.04$). These large performance differences did not persist to the final test in either experiment, and in Experiment 3a, we actually found a reversed effect for items with feedback, such that the condition with greater retrieval practice performance (4×) produced a smaller testing effect than the condition with lower retrieval practice performance (1×).

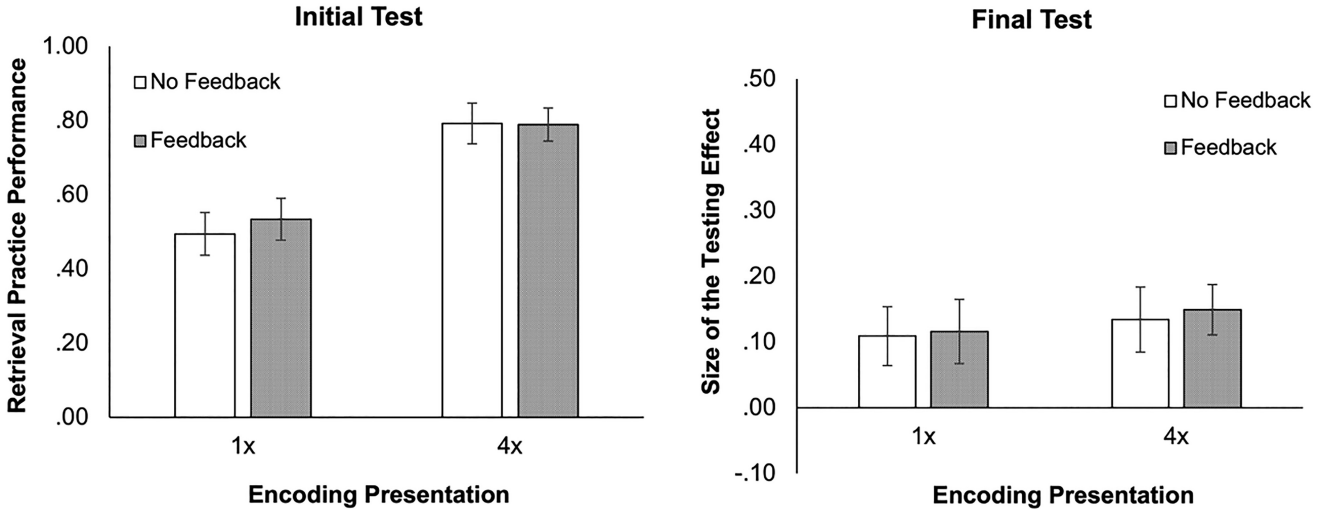
Exploratory Meta-Analysis and Correlation Analysis

Meta-Analysis of Retrieval Practice Performance and the Testing Effect

In six experiments, we showed that various manipulations that were effective at boosting retrieval practice performance had little

Figure 6

Retrieval Practice Performance and the Testing Effect as a Function of Encoding Presentation and Feedback in Experiment 3b



Note. The left and right panels had different scales for the *y* axes. Because the testing effect is a difference score between the tested and control items, it can be negative if participants recall more control items than tested items. Error bars are descriptive 0.95 confidence intervals.

impact on the eventual size of the testing effect. This conclusion was supported by very small (null) effect sizes or significant reversed effects (e.g., a larger testing effect for items recalled less frequently during the initial test). To evaluate the extent to which our data support the hypothesis that the testing effect is independent of retrieval practice performance, we also conducted Bayesian analyses. These augmented the traditional analyses and demonstrated further evidence in favor of the null (i.e., independence). However, one reservation for this conclusion may be that our experiments had relatively small sample sizes, so the null effects observed here might be the result of insufficient power to detect small effect sizes. To ameliorate this concern, we now report the results of an exploratory meta-analysis, in which we combined the data from our six experiments.

For the meta-analysis, we computed the effect size for retrieval practice recall and the testing effect by comparing the condition expected to produce the best retrieval practice performance (i.e., long response deadline in Experiment 1, two encoding trials in Experiment 2, and four encoding trials in Experiment 3) against the condition expected to produce the worst (i.e., short response deadline in Experiment 1, one encoding trial in Experiments 2 and 3). We conducted a random effects meta-analysis with the DerSimonian–Laird method to estimate the meta-analytic effect size. Further, we conducted a Bayesian model averaging meta-analysis that combined the fixed effects and random effects models (Berkhout et al., 2023; Gronau et al., 2021), which allowed us to assess support for H_0. This is particularly important when evaluating the hypothesis that retrieval practice performance and the testing effect are independent of each other. For the Bayesian meta-analysis, we set the priors of H_0 to zero and H_1 to a positively sided Cauchy distribution with $r = .707$ as the scale parameter and 0 as the lower bound. Because feedback was manipulated via different groups of participants and different sets of items (each feedback condition had its own baseline for calculating the effect size), we treated the effect size of each feedback condition as separate. Consequently, there were 12 effect sizes in this meta-analysis.

For retrieval practice performance, there was a powerful effect of our encoding- and retrieval-based manipulations, $d = 1.05$ (0.83, 1.26), $p < .001$, $BF_{+0} = 1.88 \times 10^5$, such that the condition designed to elicit better retrieval practice performance did, in fact, produce superior retrieval practice performance. Despite this large effect size on the initial test, our manipulation had virtually no influence on the size of the testing effect, $d = -0.06$ (−0.20, 0.09), $p = .463$. Indeed, the Bayesian meta-analysis strongly favored the null, $BF_{0+} = 21.58$. Figure 7 displays a forest plot for these two data sets.

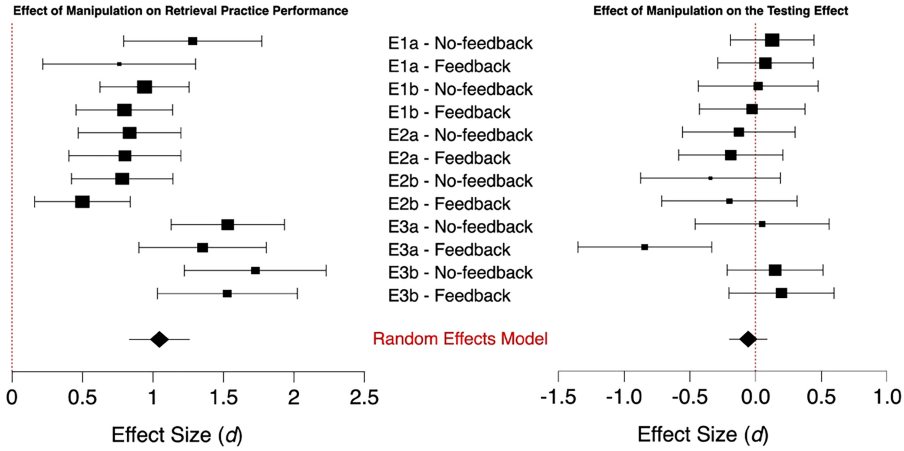
Correlation Analysis of Individual Differences

Below, we report the results of an exploratory analysis that examines whether retrieval practice performance is associated with the size of the testing effect at the individual level. Given that meta-analytic results have sometimes shown a reliable positive association between retrieval practice performance and the size of the testing effect across studies (Pan & Rickard, 2018; Rowland, 2014), our data might exhibit the same pattern across individuals. Related to the idea of individual differences, data on the relationship between learners' working memory capacity and the testing effect are mixed, with some studies showing a positive association (Tse & Pu, 2012; Zheng et al., 2023), some a null effect (Agarwal et al., 2017; Aslan & Bäuml, 2011; Pirozzolo, 2019; Storm & Bui, 2016), and some a negative association (Agarwal et al., 2017; Yang et al., 2020).[8] These conflicting findings might reflect the impact of a host of other variables inherent to their respective procedures (e.g., material type, retention interval, sample size, the specific working memory task used), and attempting to clarify these discrepancies is beyond the scope of the present study. However, because we had used RSPAN (Unsworth et al., 2005) as a distractor task in our experiments, we could examine the direction and extent

[8] Out of four regression analyses in Agarwal et al. (2017), one demonstrated a negative correlation and three showed a null effect.

Figure 7

A Forest Plot of the Effects of Our Independent Variable on Retrieval Practice Performance and the Testing Effect



Note. Left panel shows data for retrieval practice performance and right panel shows data for the testing effect, with the meta-analytic effect sizes appearing at the bottom. Effect size is in Cohen's *d*. The vertical red, dotted line shows the point at which effect size = 0. Error bars are 0.95 confidence intervals. See the online article for the color version of this figure.

to which working memory capacity was associated with the testing effect. For RSPAN, we used total number of letters recalled in the correct serial positions for analyses.

To examine whether retrieval practice performance and working memory capacity were associated with the testing effect at the individual level, we combined the data from all six experiments with the following two specifications. First, because this analysis was designed to examine individual differences in susceptibility to the testing effect, we included data from only participants or items that did not receive feedback during the initial test. Because feedback provided an additional presentation following a retrieval attempt, including the data from these items/participants would artificially weaken any putative correlation between retrieval practice performance and the testing effect. For example, imagine Subjects A and B have retrieval practice performance of .25 and .50, respectively, then their testing effects would be .10 and .20. These data would exhibit a positive correlation between retrieval practice performance and the testing effect. Now, imagine that Subjects C and D also have retrieval practice performance of .25 and .50 but received feedback; their testing effect might be .20 and .30 (because the extra presentation from the feedback increases subsequent recall relative to no-testing). When we combine the data from these four participants, the two participants with lower retrieval practice performance (Subjects A and C = .25) produce different testing effects (.10 and .20), and the same applies to the two with high retrieval practice performance. As this hypothetical example illustrates, including items/participants with feedback would contaminate the size of the testing effect and violate the aim of examining individual differences in this analysis.

Second, we included only one data point per participant. For example, in Experiment 1, because there were three response deadlines, there were three retrieval practice performances and three testing effects per participant. We opted to include only the condition with the long response deadline in Experiment 1 because this condition was the most similar to the initial test conditions under Experiments 2 and 3 (i.e., retrieval on the initial test was not constrained by a deadline). Third, because participants in Experiment 1 only had one presentation trial for each item, we included only data for items presented once in Experiments 2 and 3 (i.e., we did not include data for the 2× or 4× trials).

Based on these criteria, each participant contributed one data point for retrieval practice performance and one data point for the testing effect. The data for all participants originated in a condition with a single study trial and ample time to respond during the initial test. Consequently, should any variations arise in retrieval practice performance or the testing effect, these variations can be attributed to individual differences.

This correlation analysis included data from 204 participants, with 56 from Experiment 1, 76 from Experiment 2, and 72 from Experiment 3. Unlike the experimental results, there was a significant positive correlation between retrieval practice performance and the testing effect (see Figure 8), $r = .34$, $p < .001$, $BF_{10} = 2.07 \times 10^4$. This correlation is particularly striking when considered against the backdrop of the numerous null effects based on a manipulation of retrieval practice performance. We consider the meaning of these disparate results in the General Discussion section.

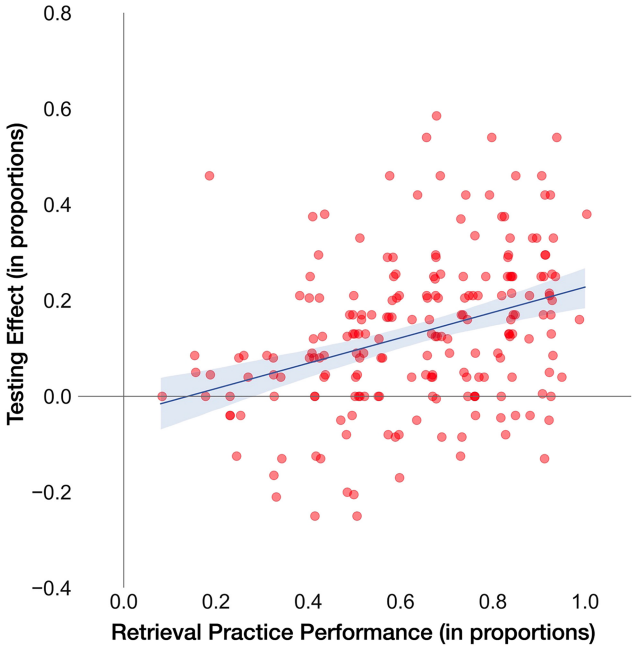
Participants' mean total RSPAN score was 52.54 ($SD = 10.71$), with a range of 13–75. The data were normally distributed with minor negative skew (−.25). Similarly, the testing effect ($M = 0.13$, $SD = 0.16$, range = −.25, .59) measure was also normally distributed (skewness = −.27). Most importantly, participants' RSPAN score did not exhibit a significant relationship with the testing effect, $r = .11$, $p = .113$, $BF_{01} = 3.28$. Notably, this null effect is observed despite the decent sample size and "well-behaved" dependent measures (i.e., normally distributed data with substantial variability).

General Discussion

In the present study, we sought to examine whether better retrieval practice performance (via an experimental manipulation) would

Figure 8

A Scatterplot Showing the Positive Correlation Between Retrieval Practice Performance and the Testing Effect Across Experiments



Note. Each dot shows the data from a single participant. Darker areas indicate greater data density. To improve data visibility, a jitter of .02 was introduced to retrieval practice performance to distribute the data horizontally. See the online article for the color version of this figure.

yield a greater testing effect. In Experiment 1, we held encoding conditions constant but varied retrieval practice performance with a response deadline. In Experiments 2 and 3, we held initial test conditions constant but varied retrieval practice performance with encoding trials. These manipulations had the desired effect on retrieval practice performance. Indeed, a meta-analysis of our data showed a robust effect size of $d = 1.03$. Despite this large increase in retrieval practice recall, we saw virtually no increase in the size of the testing effect, with the meta-analytic effect size being essentially zero ($d = -0.05$). Further, across these experiments, we showed the same pattern of results time and again, with the dissociation between retrieval practice performance and the testing effect occurring across a short or a long retention interval, and regardless of whether participants received feedback or not.

The Noncausal Relationship Between Retrieval Practice Performance and the Testing Effect

Despite our manipulations of retrieval practice performance having no impact on the testing effect, we found a significant positive correlation between retrieval practice performance and the testing effect at the individual level when participants did not receive feedback. This dissociation suggests that the naturally occurring relationship between retrieval practice performance and the testing effect is not causal; rather, it is caused by a third variable. But before we consider what this third variable could be, it is important to first address why there was a positive association between retrieval practice performance and the testing effect.

It is perhaps not controversial to suggest that participants' retrieval practice performance should be highly correlated with their final test performance for the tested items (across all six experiments of the present study, this correlation was $r = .76$). After all, participants are asked to recall the same items in both tests. We expect participants' retrieval practice performance to also correlate with their final recall of the previously nontested items, but the correlation should be weaker here because the items tested during the initial and final tests are different. Because the testing effect is a difference score between final recall of the tested items and nontested items, the strength of the correlation between the testing effect and retrieval practice performance is a function of their individual correlations. Figure 9 provides an illustration of this relationship. To provide further proof of this argument, we have developed a set of simulation data (see the Correlation Simulation VBA.xlsm file on OSF) that depict various scenarios in the Appendix. To conclude, the positive association between retrieval practice performance and the testing effect is driven entirely by individual differences.

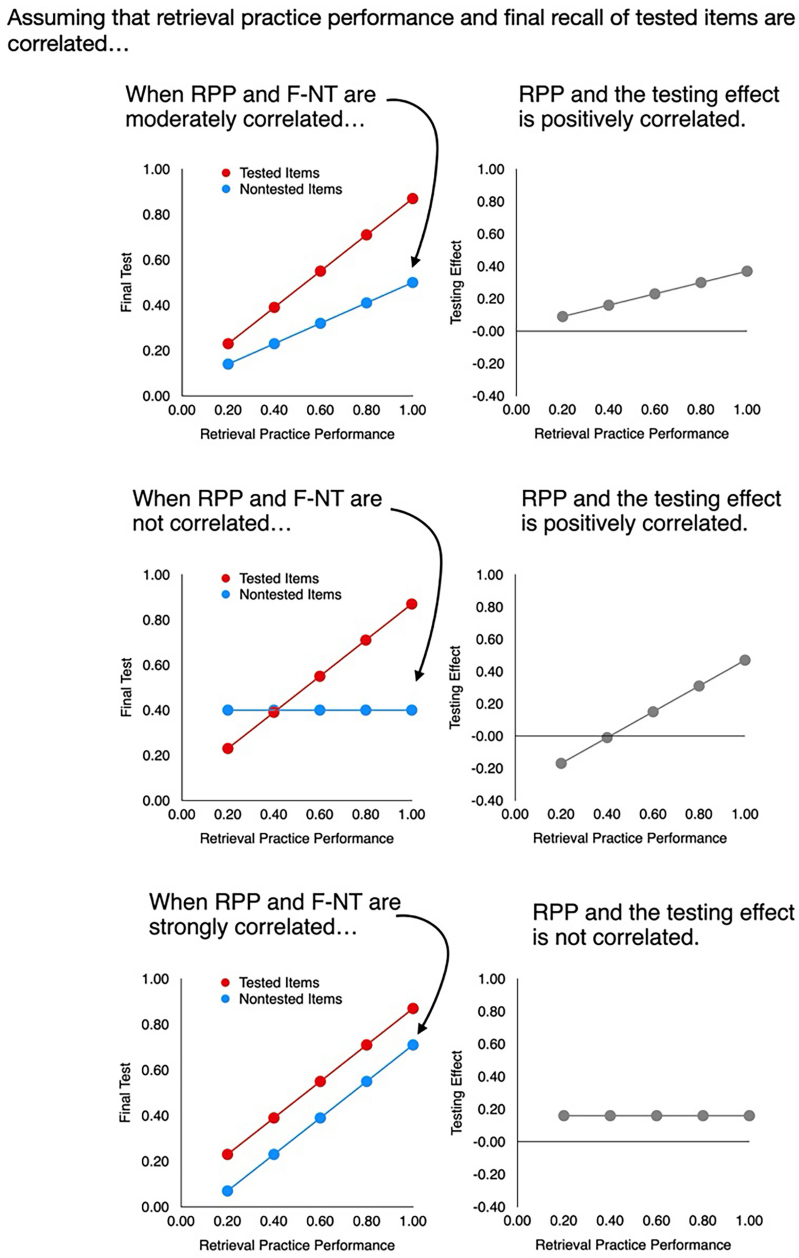
What individual difference variable(s) might contribute to this relationship? Although we caution against overspeculating here, candidate mechanisms include individual differences in learning efficiency (McDermott & Zerr, 2019), motivation (Yang et al., 2021), need for cognition (Bertilsson et al., 2017), working memory capacity (Agarwal et al., 2017; Aslan & Bäuml, 2011; Brewer & Unsworth, 2012; Storm & Bui, 2016; Yang et al., 2020), test anxiety (Tse & Pu, 2012), episodic memory ability (Brewer & Unsworth, 2012; Pan et al., 2015), self-regulated learning (Jonsson et al., 2021), intelligence (Minear et al., 2018; Robey, 2019), learning strategies (Mulligan et al., 2018; Robey, 2019), etc. In recent years, researchers have been increasingly focused on uncovering individual difference factors that might moderate learners' susceptibility to the testing effect. To date, the effort has produced mixed results (for a review, see Unsworth, 2019), so this issue awaits further research.

In addition to the independence between retrieval practice performance and the testing effect (as evidenced by our manipulations), several other noteworthy findings emerged. In Experiment 1, we found that participants could benefit from retrieval practice with a very short response deadline, and indeed, the advantage stemming from the brief retrieval practice phase was comparable to that from a retrieval practice phase that lasted more than twice as long. Recent research has begun to investigate learning strategies that are both beneficial and efficient, such as watching lectures at faster speeds (Murphy et al., 2022; Risko et al., 2023). Here, we showed that a speeded retrieval practice exercise might be as effective as a slower one. Given the importance of repeated retrieval to learning (Karpicke & Roediger, 2008), the benefits that students can attain from two speeded retrieval attempts might exceed those from a single, longer retrieval attempt. Alternatively, one might do a round of speeded retrieval and use the time saved to relearn the items (Rawson & Dunlosky, 2011). Students are often pressed for time in their busy lives, and they cite insufficient time as a primary reason against employing effective learning practices (Rea et al., 2022). Despite its important educational implication, to our knowledge, little research has examined the influence of response deadline and speeded retrieval practice on retention and transfer. Future research might benefit from a more comprehensive investigation of this practice.

Across Experiments 2 and 3, we found that differences in initial learning, which had a powerful effect on retrieval practice performance, did not affect the size of the testing effect. The null effects in Experiment 3 are especially striking given the magnitude of the effect sizes observed on retrieval practice performance ($d = 2.38$

Figure 9

Hypothetical Data Illustrating Why Retrieval Practice Performance Is Correlated With the Testing Effect on an Individual Level



Note. The right panel shows the difference between performance of the tested (the red/darker line with circles) and nontested items (the blue/lighter line with diamonds) in the corresponding left panel. RPP = retrieval practice performance; F-NT = final recall of nontested items. See the online article for the color version of this figure.

in Experiment 3a and $d = 1.97$ in Experiment 3b). Although this result might seem surprising, it dovetails with recent evidence that individuals' susceptibility to the testing effect is independent of prior knowledge (Buchin & Mulligan, 2023). Specifically, Buchin and Mulligan's high prior knowledge participants exhibited better retrieval practice performance than their low prior knowledge participants, the two groups of participants had comparable testing effects. Consequently, our findings suggest that most learners can benefit from retrieval practice (Jonsson et al., 2021).

Theoretical Implications

In the introductory part, we argued that all of the major accounts of the testing effect predict a positive relationship between retrieval

practice performance and the magnitude of the testing effect. How can these accounts reconcile with our results? Assuming that the independence of retrieval practice and the testing effect is replicable and generalizable, existing accounts might need to accommodate the idea that the processes involved when attempting to retrieve an item are as important as the success of retrieving an item. But considering how existing accounts explain the testing effect, the success of retrieval should be a critical element in achieving the benefits of testing. For example, for the episodic context account, it is difficult to reconcile how one can bind a new context to the original context of an item if the item is not retrieved (Karpicke et al., 2014). One might suggest that learners can bind the new context with the cue instead of the target, but for this explanation to accommodate our data, one must also assume that context updating for the cue alone (when retrieval practice fails) is as beneficial to subsequent memory as context updating for both the cue and the target (when retrieval practice succeeds), which seems unlikely given the logic of the framework.

For the elaborative retrieval account, it is difficult to see how one might increase retrieval routes to a target without being able to retrieve said target (Carpenter, 2009). As with the context account, one might argue that learners can increase retrieval routes from the cue (rather than to the target) even if retrieval practice fails, but this logic would again require the assumption that increasing retrieval routes to the targets is either not advantageous at all or not advantageous beyond increasing retrieval routes from the cues. In our opinion, neither of these assumptions seem consistent with the account, making it unlikely that the elaborative retrieval account as currently constructed can explain the pattern of results observed in the present studies.

Lastly, the bifurcation account (Kornell et al., 2011) does not directly explain how or why retrieval enhances retention, as it is a descriptive rather than a prescriptive account. It assumes that retrieval boosts memory strength for successfully retrieved items far more than restudy. Given this underlying assumption, the account should predict a greater testing effect under conditions when retrieval practice is more likely to succeed. Again, without extensive modification to the underlying framework of this model, the current data cannot be easily explained.

Thus far in this article, our theoretical considerations of the testing effect have not been entirely exhaustive, in part because other theoretical frameworks either do not provide straightforward predictions for the procedure used here (e.g., desirable difficulties) or they have received only minor support from previous studies (e.g., transfer appropriate processing). However, one recent notable account of the testing effect is the dual memory framework (Rickard & Pan, 2018), which proposed that the testing benefit stems from the creation of a new test memory (through successful retrieval or feedback after unsuccessful retrieval) in addition to the study memory, whereas restudying simply strengthens the existing study memory. According to this account, the additional test memory allows for more retrieval routes to the target.

On a conceptual level, there is some resemblance between this account and the context updating account or elaborative retrieval account. But this account is also accompanied by a quantitative model which allows for unambiguous hypotheses for performance in our design. Gupta et al. (2022) suggested that the model would predict a negative effect of study repetitions on the testing effect (i.e., items with more study repetitions would benefit less from retrieval practice than items with fewer study repetitions), but this hypothesis was not supported by both their results and our own. Further, it is not entirely clear whether this negative effect prediction is warranted based on the model's conceptual framework. Specifically, when learners do not receive feedback, the dual memory framework suggests that test memories can only form when retrieval success occurs, and because study repetitions increases the likelihood of retrieval practice success, it stands to reason that this model should also predict a positive effect of study repetitions on the testing effect if no-feedback is provided.

A possible solution to this conundrum is to suggest that nonretrieved targets during the initial test can still benefit from the attempt to retrieve because these targets might receive partial activation as participants attempt to retrieve the target from the cue. But an explanation like this must be thoroughly explicated and rigorously investigated, because it can be difficult to ascertain empirically whether partial activation has occurred, and what the effects of this activation are on later retrieval. Without clear empirical support, such an argument would be necessarily circular (i.e., one could argue that conditions with higher memory performance must have received higher levels of partial activation). That is beyond the scope of this article, but we encourage researchers interested in this phenomenon to explore partial activation via retrieval practice in future work.

The distinction between retrieval attempts and retrieval success might remind some readers of the two-stage framework of elaborative retrieval (Kornell et al., 2015; Kornell & Vaughn, 2016; Vaughn et al., 2017). Although this framework does consider the processes involved in a retrieval attempt (e.g., activating the semantic network associated with the retrieval cue) to be separate from those involved in retrieval success (e.g., postretrieval processing of the answer, it was designed to address a different question than ours. According to this two-stage framework, a retrieval attempt can enhance learning of the target equally regardless of whether the target is reencoded upon successful retrieval or if the target is presented as feedback to participants when retrieval fails (see also Rickard & Pan, 2018). Therefore, regarding retrieval success, the two-stage framework deals with how a target is reencoded. The present study, however, deals with whether a target is reencoded. Although this framework currently does not address our research question, it holds promise because it explicitly considers retrieval attempt and success as distinct stages. The notion of partial activation discussed above may also play a role within this framework, but such an assumption is purely speculative at this point.

The partial activation idea places the locus of our finding at the stage of retrieval practice. An alternative idea focuses on processes that occur following retrieval practice—specifically, it is possible that memories strengthened by retrieval practice show different forgetting depending on their strength or how they are recalled during the initial test. For example, stronger memories (e.g., items studied 4 times) might exhibit faster forgetting than weaker ones (e.g., items studied once). Although this suggestion seems counterintuitive, it is consistent with empirical results that memories with greater initial strength (at the starting point of a forgetting function) sometimes demonstrate more apparent forgetting than their weaker counterparts. Because weaker memories are closer to the floor of the forgetting curve (i.e., all forgetting functions are bound by zero), they would often exhibit less apparent forgetting over time than stronger memories—even if they have the same underlying forgetting rates. Interested readers may refer to several relevant papers (Bäuml, 1996; Bogartz, 1990; Loftus, 1985; Rivera-Lares et al., 2022; Rose, 1992; Slamecka, 1985; Wixted, 1990) to explore the mathematical measurement of forgetting over time.

Limitations and Constraints on Generality

One might question whether our conclusion would have differed if we had used restudy instead of no-test control as the baseline condition. We believe that it would not. Specifically, because the testing effect was assessed by a difference score between the condition(s) in which participants had completed an initial test against the condition in which they had not, substituting the no-test control condition with a restudy condition should reduce the size of the testing effect across the board (because a restudy condition would elevate performance in the baseline condition), but it would not alter the pattern of our results. Nevertheless, our conjecture here does not replace empirical evidence, so further research is needed to fully address this question.

Following the presentation of Experiments 1a and 1b, we entertained the possibility that the response deadline manipulation did not have a real impact on retrieval practice. Rather, one might argue that the shorter response deadlines merely constrained what participants could physically produce rather than what they could mentally (covertly) retrieve. This reasoning is likely incorrect. Although some studies have shown that covert retrieval practice can be as beneficial to memory retention as overt retrieval practice (Putnam & Roediger, 2013; Smith et al., 2013), more recent studies have repeatedly shown that covert retrieval practice is inferior (Jönsson et al., 2014; Kubik et al., 2020; Sumeracki & Castillo, 2022; Sundqvist et al., 2017). Further, the researchers of these studies often went to great lengths to ensure that their manipulation affected only the output format. Our response deadline manipulation, however, was designed specifically to terminate retrieval prematurely.

We also suggest that any such reasoning, upon seeing a null effect of response deadline on the testing effect, could be considered circular. But beyond that, it is well established that shorter response deadlines can have a profound impact on retrieval, particularly with effortful processes like recollection (Gardiner et al., 1999; Yonelinas, 2002). Given the putative role that recollection plays in the testing effect (Chan & McDermott, 2007; Shaffer & McDermott, 2022), it is reasonable to assume that our response deadline manipulation would have an impact on the testing effect even if we allow for the idea that actual retrieval may be underestimated by the responses collected within the shorter deadlines.

One final note of caution is that we did not implement a response deadline or an encoding trial that is extremely short (e.g., 1 s). We opted not to do this because, as we had specified in the Introduction, such a deadline would limit any retrieval occurring at all. Should this condition produce no testing effect, the result would hold little theoretical interest because floor effects are uninterpretable.

Another limitation of the present study is that we used simple rather than more complex materials (such as prose passages or video lectures) for learning. This feature of our study limits the generalizability of our conclusions somewhat. But we see no a priori reasons to suspect that our results would not also apply to more complex materials, given that reviews of the literature have consistently found that retrieval practice enhances learning across item complexities (Adesope et al., 2017; Rowland, 2014; Yang et al., 2021). But one potentially important issue bears mentioning: our experiments used weakly associated word pairs as learning materials, which necessarily means that we were examining associative memory. From this perspective, one might wonder whether our results would generalize to arbitrary associations (i.e., unrelated pairs), because other educationally relevant phenomena have sometimes demonstrated dissociations for related and unrelated pairs (e.g., Grimaldi & Karpicke, 2012; but see also Potts & Shanks, 2014). To partially address this possibility, we conducted an exploratory analysis of our data in Experiments 2 and 3. Specifically, participants studied 96 weakly associated word pairs in these experiments. Although all of these pairs had an identical, weak backward association ($M_{BAS} = 0.05$ for every pair), they differed in their forward associative strength ($M_{FAS} = 0.01$, $SD = 0.02$). Critically, 57 of the pairs had a zero FAS, and the remaining 39 pairs had a positive one ($M_{FAS} = 0.03$, $SD = 0.02$). Consequently, the former set of items were, by definition, unrelated in the cue-to-target direction (which was how we tested participants' memory). Although none of our pairs were truly unrelated, this analysis provided at least an opportunity to examine whether our main finding held across degrees of relatedness.

To this end, we combined the data from Experiments 2 and 3 and conducted a repeated measures ANOVA with relatedness (i.e., whether the word pairs were related or unrelated in FAS), feedback, and study trials (once vs. multiple) serving as independent variables, and the testing effect magnitude serving as the dependent variable. Relatedness did not produce any significant interactions, all $ps > .433$, $BF_{01}s > 6.06$. For present purposes, the most important finding is that both unrelated ($M_{once} = 0.18$, $M_{multiple} = 0.18$) and related items ($M_{once} = 0.17$, $M_{multiple} = 0.14$) yielded no increase in the testing effect with study repetitions. Indeed, relatedness did not produce an interaction regardless of whether we examined the data of Experiments 2 and 3 in aggregate or individually. Therefore, the independence between retrieval practice performance and the testing effect applies equally well to both sets of items in these experiments.

Finally, although we have conducted six experiments and shown the same results when employing a manipulation that affected encoding (number of study trials), retrieval practice (response deadline), postretrieval processing (feedback), and storage (retention interval), these data still came from a single university, and we have only conducted within-subjects experiments with paired associates. There are many more possible manipulations that one can examine (e.g., different participant populations, repeated retrieval practice, various material types, including examining item memory rather than associative memory), so we urge caution when attempting to draw broad conclusions from the present data.

Conclusion

In six experiments, we showed that the magnitude of the testing effect is independent of retrieval practice performance when the latter was manipulated experimentally. This independence was observed under both an encoding- and retrieval-based manipulation, at both a short and long retention interval, and was unaffected by feedback. This same conclusion was supported by a more highly powered meta-analysis of our combined data. In addition, an exploratory analysis revealed a significant positive association between retrieval practice performance and the testing effect at an individual level. The dissociation between the results of our experimental manipulations and the naturally occurring association suggests that the positive correlation is caused by a third variable and not by retrieval practice success or failure. Together, the present data highlighted a crucial misconception in the current understanding of the testing effect, both at the empirical and theoretical levels. Further, the current data support the idea that retrieval practice is beneficial to learners even when the initial test

conditions are not conducive to success (e.g., few encoding attempts or a short retrieval window). This finding further validates the approach of implementing retrieval practice "early and often (Yan et al., 2016)" rather than waiting until learners can answer a majority of questions correctly, which may encourage more educators and students to apply this strategy in real-world contexts.

References

Adesope, O. O., Trevisan, D. A., & Sundararajan, N. (2017). Rethinking the use of tests: A meta-analysis of practice testing. *Review of Educational Research*, *87*(3), 659–701. https://doi.org/10.3102/0034654316689306

Agarwal, P. K., Finley, J. R., Rose, N. S., & Roediger, H. L. (2017). Benefits from retrieval practice are greater for students with lower working memory capacity. *Memory*, *25*(6), 764–771. https://doi.org/10.1080/09658211.2016.1220579

Agarwal, P. K., Karpicke, J. D., Kang, S. H. K., Roediger, H. L., & McDermott, K. B. (2008). Examining the testing effect with open- and closed-book tests. *Applied Cognitive Psychology*, *22*(7), 861–876. https://doi.org/10.1002/acp.1391

Ahn, D., & Chan, J. C. K. (2022). Does testing enhance new learning because it insulates against proactive interference? *Memory & Cognition*, *50*(8), 1664–1682. https://doi.org/10.3758/s13421-022-01273-7

Ahn, D., & Chan, J. C. K. (2023). Does testing potentiate new learning because it enables learners to use better strategies? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *50*(3), 435–457. https://doi.org/10.1037/xlm0001233

Appelbaum, M., Cooper, H., Kline, R. B., Mayo-Wilson, E., Nezu, A. M., & Rao, S. M. (2018). Journal article reporting standards for quantitative research in psychology: The APA Publications and Communications Board task force report. *American Psychologist*, *73*(1), 3–25. https://doi.org/10.1037/amp0000191

Arnold, K. M., & McDermott, K. B. (2013a). Free recall enhances subsequent learning. *Psychonomic Bulletin & Review*, *20*(3), 507–513. https://doi.org/10.3758/s13423-012-0370-3

Arnold, K. M., & McDermott, K. B. (2013b). Test-potentiated learning: Distinguishing between direct and indirect effects of tests. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *39*(3), 940–945. https://doi.org/10.1037/a0029199

Aslan, A., & Bäuml, K. H. (2011). Individual differences in working memory capacity predict retrieval-induced forgetting. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *37*(1), 264–269. https://doi.org/10.1037/a0021324

Balota, D. A., & Neely, J. H. (1980). Test-expectancy and word-frequency effects in recall and recognition. *Journal of Experimental Psychology: Human Learning and Memory*, *6*(5), 576–587. https://doi.org/10.1037/0278-7393.6.5.576

Bäuml, K.-H. (1996). Revisiting an old issue: Retroactive interference as a function of the degree of original and interpolated learning. *Psychonomic Bulletin & Review*, *3*(3), 380–384. https://doi.org/10.3758/BF03210765

Berkhout, S. W., Haaf, J. M., Gronau, Q. F., Heck, D. W., & Wagenmakers, E. J. (2023). A tutorial on Bayesian model-averaged meta-analysis in JASP. *Behavior Research Methods*, *56*(3), 1260–1282. https://doi.org/10.3758/s13428-023-02093-6

Bertilsson, F., Wiklund-Hörnqvist, C., Stenlund, T., & Jonsson, B. (2017). The testing effect and its relation to working memory capacity and personality characteristics. *Journal of Cognitive Education and Psychology*, *16*(3), 241–259. https://doi.org/10.1891/1945-8959.16.3.241

Blumen, H. M., & Rajaram, S. (2008). Influence of re-exposure and retrieval disruption during group collaboration on later individual recall. *Memory*, *16*(3), 231–244. https://doi.org/10.1080/09658210701804495

Bogartz, R. S. (1990). Learning-forgetting rate independence defined by forgetting function parameters or forgetting function form: Reply to Loftus and Bamber and to Wixted. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *16*(5), 936–945. https://doi.org/10.1037/0278-7393.16.5.936

Brewer, G. A., & Unsworth, N. (2012). Individual differences in the effects of retrieval from long-term memory. *Journal of Memory and Language*, *66*(3), 407–415. https://doi.org/10.1016/j.jml.2011.12.009

Brysbaert, M. (2019). How many participants do we have to include in properly powered experiments? A tutorial of power analysis with reference tables. *Journal of Cognition*, *2*(1), Article 16. https://doi.org/10.5334/joc.72

Buchin, Z. L., & Mulligan, N. W. (2023). Retrieval-based learning and prior knowledge. *Journal of Educational Psychology*, *115*(1), 22–35. https://doi.org/10.1037/edu0000773

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. (2006). Impoverished cue support enhances subsequent retention: Support for the elaborative retrieval explanation of the testing effect. *Memory & Cognition*, *34*(2), 268–276. https://doi.org/10.3758/BF03193405

Chan, J. C. K. (2007). *An investigation of retrieval-induced facilitation* [Doctoral dissertation, Washington University in St. Louis]. ProQuest Dissertation and Theses Global.

Chan, J. C. K. (2009). When does retrieval induce forgetting and when does it induce facilitation? Implications for retrieval inhibition, testing effect, and text processing. *Journal of Memory and Language*, *61*(2), 153–170. https://doi.org/10.1016/j.jml.2009.04.004

Chan, J. C. K., Manley, K. D., & Ahn, D. (2020). Does retrieval potentiate new learning when retrieval stops but new learning continues? *Journal of Memory and Language*, *115*, Article 104150. https://doi.org/10.1016/j.jml.2020.104150

Chan, J. C. K., Manley, K. D., Davis, S. D., & Szpunar, K. K. (2018). Testing potentiates new learning across a retention interval and a lag: A strategy change perspective. *Journal of Memory and Language*, *102*, 83–96. https://doi.org/10.1016/j.jml.2018.05.007

Chan, J. C. K., & McDermott, K. B. (2007). The testing effect in recognition memory: A dual process account. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *33*(2), 431–437. https://doi.org/10.1037/0278-7393.33.2.431

Chan, J. C. K., McDermott, K. B., & Roediger, H. L. (2006). Retrieval-induced facilitation: Initially nontested material can benefit from prior testing of related material. *Journal of Experimental Psychology: General*, *135*(4), 553–571. https://doi.org/10.1037/0096-3445.135.4.553

Chan, J. C. K., Meissner, C. A., & Davis, S. D. (2018). Retrieval potentiates new learning: A theoretical and meta-analytic review. *Psychological Bulletin*, *144*(11), 1111–1146. https://doi.org/10.1037/bul0000166

Cho, K. W., & Neely, J. H. (2017). The roles of encoding strategies and retrieval practice in test-expectancy effects. *Memory*, *25*(5), 626–635. https://doi.org/10.1080/09658211.2016.1202983

Cranney, J., Ahn, M., McKinnon, R., Morris, S., & Watts, K. (2009). The testing effect, collaborative learning, and retrieval-induced facilitation in a classroom setting. *European Journal of Cognitive Psychology*, *21*(6), 919–940. https://doi.org/10.1080/09541440802413505

de Lima, M. F. R., Venâncio, S., Feminella, J., & Buratto, L. G. (2020). Does item difficulty affect the magnitude of the retrieval practice effect? An evaluation of the retrieval effort hypothesis. *The Spanish Journal of Psychology*, *23*, Article e31. https://doi.org/10.1017/SJP.2020.33

Duchastel, P. C. (1981). Retention of prose following testing with different types of tests. *Contemporary Educational Psychology*, *6*(3), 217–226. https://doi.org/10.1016/0361-476X(81)90002-3

Fiechter, J. L., & Benjamin, A. S. (2018). Diminishing-cues retrieval practice: A memory-enhancing technique that works when regular testing doesn't. *Psychonomic Bulletin & Review*, *25*(5), 1868–1876. https://doi.org/10.3758/s13423-017-1366-9

Gardiner, J. M., Ramponi, C., & Richardson-Klavehn, A. (1999). Response deadline and subjective awareness in recognition memory. *Consciousness and Cognition*, *8*(4), 484–496. https://doi.org/10.1006/ccog.1999.0409

Glover, J. A. (1989). The "testing" phenomenon: Not gone but nearly forgotten. *Journal of Educational Psychology*, *81*(3), 392–399. https://doi.org/10.1037/0022-0663.81.3.392

Grimaldi, P. J., & Karpicke, J. D. (2012). When and why do retrieval attempts enhance subsequent encoding? *Memory & Cognition*, *40*(4), 505–513. https://doi.org/10.3758/s13421-011-0174-0

Gronau, Q. F., Heck, D. W., Berkhout, S. W., Haaf, J. M., & Wagenmakers, E.-J. (2021). A primer on Bayesian model-averaged meta-analysis. *Advances in Methods and Practices in Psychological Science*, *4*(3), Article 251524592110312. https://doi.org/10.1177/25152459211031256

Gupta, M. W., Pan, S. C., & Rickard, T. C. (2022). Prior episodic learning and the efficacy of retrieval practice. *Memory & Cognition*, *50*(4), 722–735. https://doi.org/10.3758/s13421-021-01236-4

Indig, S. (2005). *Cramming Canadians: Two-thirds of university students start studying for exams no more than a week in advance*. kumon.com. https://www.kumon.com/pressroom/pressreleases/article-canada.asp?articlenum=17&language=Canada

Izawa, C. (1971). The test trial potentiating model. *Journal of Mathematical Psychology*, *8*(2), 200–224. https://doi.org/10.1016/0022-2496(71)90012-5

JASP Team. (2023). *JASP* (Version 0.18.2) [Computer software]. https://jasp-stats.org/

Jonsson, B., Wiklund-Hörnqvist, C., Stenlund, T., Andersson, M., & Nyberg, L. (2021). A learning method for all: The testing effect is independent of cognitive ability. *Journal of Educational Psychology*, *113*(5), 972–985. https://doi.org/10.1037/edu0000627

Jönsson, F. U., Kubik, V., Sundqvist, M. L., Todorov, I., & Jonsson, B. (2014). How crucial is the response format for the testing effect? *Psychological Research*, *78*(5), 623–633. https://doi.org/10.1007/s00426-013-0522-8

Kanayama, K., & Kasahara, K. (2018). The indirect effects of testing: Can poor performance in a vocabulary quiz lead to long-term L2 vocabulary retention. *Vocabulary Learning and Instruction*, *7*(1), 1–13. https://doi.org/10.7820/vli.v07.1.kanayama.kasahara

Kang, S. H. K., McDermott, K. B., & Roediger, H. L. (2007). Test format and corrective feedback modify the effect of testing on long-term retention. *European Journal of Cognitive Psychology*, *19*(4–5), 528–558. https://doi.org/10.1080/09541440601056620

Karpicke, J. D., Lehman, M., & Aue, W. R. (2014). Retrieval-based learning: An episodic context account. *The Psychology of Learning and Motivation*, *61*, 237–284. https://doi.org/10.1016/B978-0-12-800283-4.00007-1

Karpicke, J. D., & Roediger, H. L. (2008). The critical importance of retrieval for learning. *Science*, *319*(5865), 966–968. https://doi.org/10.1126/science.1152408

Kliegl, O., Bjork, R. A., & Bäuml, K.-H. T. (2019). Feedback at test can reverse the retrieval-effort effect. *Frontiers in Psychology*, *10*, Article 1863. https://doi.org/10.3389/fpsyg.2019.01863

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

Kornell, N., Hays, M. J., & Bjork, R. A. (2009). Unsuccessful retrieval attempts enhance subsequent learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *35*(4), 989–998. https://doi.org/10.1037/a0015729

Kornell, N., Klein, P. J., & Rawson, K. A. (2015). Retrieval attempts enhance learning, but retrieval success (versus failure) does not matter. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *41*(1), 283–294. https://doi.org/10.1037/a0037850

Kornell, N., & Vaughn, K. E. (2016). How retrieval attempts affect learning. In B. H. Ross (Ed.), *Psychology of learning and motivation* (pp. 183–215). Elsevier. https://doi.org/10.1016/bs.plm.2016.03.003

Kubik, V., Jönsson, F. U., de Jonge, M., & Arshamian, A. (2020). Putting action into testing: Enacted retrieval benefits long-term retention more than covert retrieval. *Quarterly Journal of Experimental Psychology*, *73*(12), 2093–2105. https://doi.org/10.1177/1747021820945560

Kuo, T. M., & Hirshman, E. (1996). Investigations of the testing effect. *American Journal of Psychology*, *109*(3), 451–464. https://doi.org/10.2307/1423016

Loftus, G. R. (1985). Evaluating forgetting curves. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *11*(2), 397–406. https://doi.org/10.1037/0278-7393.11.2.397

McDermott, K. B., & Zerr, C. L. (2019). Individual differences in learning efficiency. *Current Directions in Psychological Science*, *28*(6), 607–613. https://doi.org/10.1177/0963721419869005

Minear, M., Coane, J. H., Boland, S. C., Cooney, L. H., & Albat, M. (2018). The benefits of retrieval practice depend on item difficulty and intelligence. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *44*(9), 1474–1486. https://doi.org/10.1037/xlm0000486

Mulligan, N. W., Buchin, Z. L., & Zhang, A. L. (2022). The testing effect with free recall: Organization, attention, and order effects. *Journal of Memory and Language*, *125*, Article 104333. https://doi.org/10.1016/j.jml.2022.104333

Mulligan, N. W., Rawson, K. A., Peterson, D. J., & Wissman, K. T. (2018). The replicability of the negative testing effect: Differences across participant populations. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *44*(5), 752–763. https://doi.org/10.1037/xlm0000490

Murphy, D. H., Hoover, K. M., Agadzhanyan, K., Kuehn, J. C., & Castel, A. D. (2022). Learning in double time: The effect of lecture video speed on immediate and delayed comprehension. *Applied Cognitive Psychology*, *36*(1), 69–82. https://doi.org/10.1002/acp.3899

Nickerson, C. A., & Brown, N. J. L. (2019). Simpson's Paradox is suppression, but Lord's Paradox is neither: Clarification of and correction to Tu, Gunnell, and Gilthorpe (2008). *Emerging Themes in Epidemiology*, *16*(1), Article 5. https://doi.org/10.1186/s12982-019-0087-0

Overall, J. E., & Woodward, J. A. (1975). Unreliability of difference scores: A paradox for measurement of change. *Psychological Bulletin*, *82*(1), 85–86. https://doi.org/10.1037/h0076158

Pan, S. C., Pashler, H., Potter, Z. E., & Rickard, T. C. (2015). Testing enhances learning across a range of episodic memory abilities. *Journal of Memory and Language*, *83*, 53–61. https://doi.org/10.1016/j.jml.2015.04.001

Pan, S. C., & Rickard, T. C. (2018). Transfer of test-enhanced learning: Meta-analytic review and synthesis. *Psychological Bulletin*, *144*(7), 710–756. https://doi.org/10.1037/bul0000151

Pirozzolo, J. W. (2019). *The testing effect, individual differences, and transfer: An investigation of learning strategies using educational materials* [University of Houston]. ProQuest Dissertation and Theses Global.

Potts, R., & Shanks, D. R. (2014). The benefit of generating errors during learning. *Journal of Experimental Psychology: General*, *143*(2), 644–667. https://doi.org/10.1037/a0033194

Putnam, A. L., & Roediger, H. L. (2013). Does response mode affect amount recalled or the magnitude of the testing effect? *Memory & Cognition*, *41*(1), 36–48. https://doi.org/10.3758/s13421-012-0245-x

Racsmány, M., Szőllősi, Á, & Marián, M. (2020). Reversing the testing effect by feedback is a matter of performance criterion at practice. *Memory & Cognition*, *48*(7), 1161–1170. https://doi.org/10.3758/s13421-020-01041-5

Rawson, K. A., & Dunlosky, J. (2011). Optimizing schedules of retrieval practice for durable and efficient learning: How much is enough? *Journal of Experimental Psychology: General*, *140*(3), 283–302. https://doi.org/10.1037/a0023956

Rea, S. D., Wang, L., Muenks, K., & Yan, V. X. (2022). Students can (mostly) recognize effective learning, so why do they not do it? *Journal of Intelligence*, *10*(4), Article 127. https://doi.org/10.3390/jintelligence10040127

Rickard, T. C., & Pan, S. C. (2018). A dual memory theory of the testing effect. *Psychonomic Bulletin & Review*, *25*(3), 847–869. https://doi.org/10.3758/s13423-017-1298-4

Risko, E. F., Liu, J., & Bianchi, L. (2023). Speeding lectures to make time for retrieval practice: Can we improve the efficiency of interpolated testing?

Journal of Experimental Psychology: Applied. Advance online publication. https://doi.org/10.1037/xap0000494

Rivera-Lares, K., Logie, R., Baddeley, A., & Della Sala, S. (2022). Rate of forgetting is independent of initial degree of learning. *Memory & Cognition*, *50*(8), 1706–1718. https://doi.org/10.3758/s13421-021-01271-1

Robey, A. (2019). The benefits of testing: Individual differences based on student factors. *Journal of Memory and Language*, *108*, Article 104029. https://doi.org/10.1016/j.jml.2019.104029

Roediger, H. L., & Karpicke, J. D. (2006). 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

Rose, R. J. (1992). Degree of learning, interpolated tests, and rate of forgetting. *Memory & Cognition*, *20*(6), 621–632. https://doi.org/10.3758/bf03202712

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

Shaffer, R. A., & McDermott, K. B. (2022). The dual-process perspective and the benefits of retrieval practice in younger and older adults. *Memory*, *30*(5), 554–572. https://doi.org/10.1080/09658211.2022.2027986

Shimizu, Y., & Jacoby, L. L. (2005). Similarity-guided depth of retrieval: Constraining at the front end. *Canadian Journal of Experimental Psychology*, *59*(1), 17–21. https://doi.org/10.1037/h0087455

Slamecka, N. J. (1985). On comparing rates of forgetting: Comment on Loftus (1985). *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *11*(4), 812–816. https://doi.org/10.1037/0278-7393.11.1-4.812

Smith, M. A., & Karpicke, J. D. (2014). Retrieval practice with short-answer, multiple-choice, and hybrid tests. *Memory*, *22*(7), 784–802. https://doi.org/10.1080/09658211.2013.831454

Smith, M. A., Roediger, H. L., & Karpicke, J. D. (2013). Covert retrieval practice benefits retention as much as overt retrieval practice. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *39*(6), 1712–1725. https://doi.org/10.1037/a0033569

St. Hilaire, K. J., Chan, J. C. K., & Ahn, D. (2023). Guessing as a learning intervention: A meta-analytic review of the prequestion effect. *Psychonomic Bulletin & Review*. Advance online publication. https://doi.org/10.3758/s13423-023-02353-8

Storm, B. C., & Bui, D. C. (2016). Retrieval-practice task affects relationship between working memory capacity and retrieval-induced forgetting. *Memory*, *24*(10), 1407–1418. https://doi.org/10.1080/09658211.2015.1117640

Sumeracki, M. A., & Castillo, J. (2022). Covert and overt retrieval practice in the classroom. *Translational Issues in Psychological Science*, *8*(2), 282–293. https://doi.org/10.1037/tps0000332

Sundqvist, M. L., Mäntylä, T., & Jönsson, F. U. (2017). Assessing boundary conditions of the testing effect: On the relative efficacy of covert versus overt retrieval. *Frontiers in Psychology*, *8*, Article 1018. https://doi.org/10.3389/fpsyg.2017.01018

Thomas, D. R., & Zumbo, B. D. (2011). Difference scores from the point of view of reliability and repeated-measures ANOVA. *Educational and Psychological Measurement*, *72*(1), 37–43. https://doi.org/10.1177/0013164411409929

Thompson, C. P., Wenger, S. K., & Bartling, C. A. (1978). How recall facilitates subsequent recall: A reappraisal. *Journal of Experimental Psychology: Human Learning and Memory*, *4*(3), 210–221. https://doi.org/10.1037/0278-7393.4.3.210

Toppino, T. C., & Cohen, M. S. (2009). The testing effect and the retention interval. *Experimental Psychology*, *56*(4), 252–257. https://doi.org/10.1027/1618-3169.56.4.252

Trafimow, D. (2015). A defense against the alleged unreliability of difference scores. *Cogent Mathematics*, *2*(1), Article 1064626. https://doi.org/10.1080/23311835.2015.1064626

Tse, C.-S., & Pu, X. (2012). The effectiveness of test-enhanced learning depends on trait test anxiety and working-memory capacity. *Journal of Experimental Psychology: Applied*, *18*(3), 253–264. https://doi.org/10.1037/a0029190

Unsworth, N. (2019). Individual differences in long-term memory. *Psychological Bulletin*, *145*(1), 79–139. https://doi.org/10.1037/bul0000176

Unsworth, N., Heitz, R. P., Schrock, J. C., & Engle, R. W. (2005). An automated version of the operation span task. *Behavior Research Methods*, *37*(3), 498–505. https://doi.org/10.3758/BF03192720

van Ravenzwaaij, D., & Wagenmakers, E. J. (2022). Advantages masquerading as "issues" in Bayesian hypothesis testing: A commentary on Tendeiro and Kiers (2019). *Psychological Methods*, *27*(3), 451–465. https://doi.org/10.1037/met0000415

Vaughn, K. E., Hausman, H., & Kornell, N. (2017). Retrieval attempts enhance learning regardless of time spent trying to retrieve. *Memory*, *25*(3), 298–316. https://doi.org/10.1080/09658211.2016.1170152

Vojdanoska, M., Cranney, J., & Newell, B. R. (2009). The testing effect: The role of feedback and collaboration in a tertiary classroom setting. *Applied Cognitive Psychology*, *24*(8), 1183–1195. https://doi.org/10.1002/acp.1630

Whitten, W. B. (1978). Initial-retrieval "depth" and the negative recency effect. *Memory & Cognition*, *6*(6), 590–598. https://doi.org/10.3758/BF03198248

Wixted, J. T. (1990). Analyzing the empirical course of forgetting. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *16*(5), 927–935. https://doi.org/10.1037/0278-7393.16.5.927

Yan, V. X., Clark, C. M., & Bjork, R. A. (2016). Memory and metamemory considerations in the instruction of human beings revisited: Implications for optimizing online learning. In J. C. Horvath, J. M. Lodge, & J. Hattie (Eds.), *From the laboratory to the classroom: Translating science of learning for teachers* (pp. 61–78). Routledge. https://books.google.com/books?hl=en&lr=&id=qjK6DAAAQBAJ&oi=fnd&pg=PP1&dq=10.4324/9781315625737&ots=xn3rgmTgX3&sig=Jv1MZgGqV5R_aaG3qo9ZokTLxCU

Yang, C., Luo, L., Vadillo, M. A., Yu, R., & Shanks, D. R. (2021). Testing (quizzing) boosts classroom learning: A systematic and meta-analytic review. *Psychological Bulletin*, *147*(4), 399–435. https://doi.org/10.1037/bul0000309

Yang, C., Sun, B., Potts, R., Yu, R., Luo, L., & Shanks, D. R. (2020). Do working memory capacity and test anxiety modulate the beneficial effects of testing on new learning? *Journal of Experimental Psychology: Applied*, *26*(4), 724–738. https://doi.org/10.1037/xap0000278

Yonelinas, A. P. (2002). The nature of recollection and familiarity: A review of 30 years of research. *Journal of Memory and Language*, *46*(3), 441–517. https://doi.org/10.1006/jmla.2002.2864

Yurtsever, A., Chan, J. C. K., Davis, S. D., & Myers, S. J. (2024). *Testing effect and retrieval practice*. Open Science Framework. https://osf.io/st596/

Zheng, Y., Sun, P., & Liu, X. L. (2023). Retrieval practice is costly and is beneficial only when working memory capacity is abundant. *npj Science of Learning*, *8*(1), Article 8. https://doi.org/10.1038/s41539-023-00159-w

(*Appendix follows*)

Appendix

Data Simulation

The simulated data were constructed using two basic assumptions: First, we assumed that retrieval practice performance (RPP) and final recall of the tested items (F-T) are strongly correlated—for the simulation, we set this correlation to about .70, which corresponds with our empirical data. Second, we assumed that F-T would, on average, exceed final recall of the nontested items (F-NT), but both would exhibit some forgetting relative to RPP (because of the retention interval). For the simulation, we held standard deviations similar for all final recall probabilities, and we set F-NT to be about .15 worse than RPP, and F-T to be about .05 worse than RPP (based on the assumption that retrieval practice reduces forgetting). We then created three separate simulation scenarios for F-NT. Across the three scenarios, F-NT is was either uncorrelated with RPP, moderately correlated with RPP (at about .30), or strongly correlated with RPP (at about .70). We believe that, under most situations, the moderately correlated scenario is the most likely, given persistent individual differences in memory performance (i.e., some participants have a better episodic memory than others).

The results of the simulations are shown in Table A1. Each simulation comprised data from 200 virtual participants and was run 1,000 times. Interested readers can download the simulation file from our OSF page and easily run their own simulations by executing the embedded VBA script. The formulae for generating RPP, F-T, and F-NT are also editable. Note that the exact numbers in the first data column of Table A1 are not important, as the simulations were specifically designed to produce uncorrelated, moderately correlated, and strongly correlated data. As can be seen, the correlation between RPP and the testing effect, as long as RPP and F-T are more strongly associated than RPP and F-NT, would be inversely related to the correlation between RPP and F-NT. Indeed, only when the RPP × F-NT correlation and RPP × F-T correlation become very similar would one observe an independence between RPP and the testing effect.

Table A1
Results of 1,000 Simulations When RPP and F-T Are Assumed to Be Strongly Correlated

Item type	Mean proportion correct (*SD*)	*r* (*SD*)	Range	Correlation between RPP and testing effect (*SD*)
RPP	0.51 (0.02)			
F-T	0.43 (0.02)			
F-NT (not correlated with RPP)	0.35 (0.02)			
F-NT (moderately correlated with RPP)	0.36 (0.02)			
F-NT (strongly correlated with RPP)	0.35 (0.02)			
RPP × F-T Strongly correlated		.67 (0.04)	.55, .76	
RPP × F-NT Not correlated		.00 (0.07)	−.24, .22	.48 (0.05)
RPP × F-NT Moderately correlated		.35 (0.06)	.12, .54	.23 (0.07)
RPP × F-NT Strongly correlated		.67 (0.04)	.49, .76	.02 (0.07)

Note. The rightmost column shows the correlation between retrieval practice performance and the magnitude of the testing effect (i.e., F-T minus F-NT). RPP and F-T were held constant across all three scenarios (see the means in the top two cells in the leftmost data column), and F-NT varied by the level of correlations between RPP and F-NT (i.e., not correlated, moderately correlated, and strongly correlated). RPP and the testing effect is an artifact necessitated by individual differences in memory, rather than something that requires a cognitive explanation. RPP = retrieval practice performance; F-T = final recall of initially tested items; F-NT = final recall of nontested items.

Received October 17, 2023
Revision received February 22, 2024
Accepted March 11, 2024 ■