Pretesting Versus Posttesting: Comparing the Pedagogical Benefits of Errorful Generation and Retrieval Practice

Steven C. Pan[1] and Faria Sana[2]

[1]Department of Psychology, University of California, Los Angeles

[2]Centre for Social Sciences, Athabasca University

Word count (main text): 12,516

This manuscript was accepted for publication in the *Journal of Experimental Psychology: Applied* on November 12, 2020. This document may not exactly replicate the final version published in the APA journal. It is not the copy of record. The final version will be available via the journal website at: dx.doi.org/10.1037/xap0000345.

Author Note

Steven C. Pan https://orcid.org/0000-0001-9080-5651

Faria Sana http://orcid.org/0000-0002-2202-7592

Both authors contributed equally to this work.

We have no known conflict of interest to disclose. Thanks to Jeri Little for providing the stimulus materials, Elizabeth Ligon Bjork and Robert Bjork for helpful comments during the course of this research, Tim Rickard and anonymous reviewers for feedback on an earlier draft of this manuscript, Yunning Qiu for assistance with data processing and meta-analysis, and members of the Bjork Learning and Forgetting Laboratory that assisted with the running of Experiment 4.

Correspondence regarding this manuscript should be addressed to Steven C. Pan at the Department of Psychology, UCLA (Mailing address: 502 Portola Plaza, 1285 Pritzker Hall, Box 951563, Los Angeles, CA, USA, 90095-1563; E-mail: stevencpan@psych.ucla.edu), or Faria Sana at the Centre for Social Sciences, Athabasca University (Mailing address: 10011 109 St NW #1200, Edmonton, Alberta, Canada T5J 3S8; E-mail: fsana@athabascau.ca).

Data and materials for this study are archived at the Open Science Framework at: https://osf.io/zd86x/

Abstract

The use of practice tests to enhance learning, or *test-enhanced learning*, ranks among the most effective of all pedagogical techniques. We investigated the relative efficacy of *pretesting* (i.e., errorful generation) and *posttesting* (i.e., retrieval practice), two of the most prominent practice test types in the literature to date. Pretesting involves taking tests *before* to-be-learned information is studied, whereas posttesting involves taking tests *after* information is studied. In five experiments (combined n = 1,573), participants studied expository text passages, each paired with a pretest or a posttest. The tests involved multiple-choice (Experiments 1-5) or cued recall format (Experiments 2-4) and were administered with or without correct answer feedback (Experiments 3-4). On a criterial test administered 5 minutes or 48 hours later, both test types enhanced memory relative to a no-test control, but pretesting yielded higher overall scores. That advantage held across test formats, in the presence or absence of feedback, at different retention intervals, and appeared to stem from enhanced processing of text passage content (Experiment 5). Thus, although the benefits of posttesting are more well-established in the literature, pretesting is highly competitive with posttesting and can yield similar, if not greater, pedagogical benefits. These findings have important implications for the incorporation of practice tests in education and training contexts.

Keywords: pretesting; posttesting; prequestions, retrieval practice, forward and backward testing effect, test-potentiated learning

Public Significance Statement

The present study reveals that taking a practice test before a text passage is read (*pretesting*) can yield similar and, in a variety of circumstances, greater learning benefits than taking a practice test after a text passage is read (*posttesting*). Both types of tests improve memory for tested information and sometimes also improve memory for untested information. Thus, posttesting is not the only viable form of practice testing; both methods can be beneficial for learning.

Pretesting Versus Posttesting: Comparing the Pedagogical Benefits of Errorful Generation and Retrieval Practice

Although more commonly used for assessment, tests can also function as potent learning devices. Test-taking can improve memory (Rowland, 2014), increase transfer of learning in various situations (Pan & Rickard, 2018), and enhance the encoding of new information (Chan et al., 2018). The pedagogical benefits of testing, or *test-enhanced learning*, are more likely when tests are low-stakes rather than high-stakes (Hinze & Rapp, 2014), as in the case of practice quizzes as opposed to graded exams. Test-enhanced learning has been demonstrated across different learning materials (Dunlosky et al., 2013; for a list, see Rawson & Dunlosky, 2011), with diverse learners (e.g., Meyer & Logan, 2013), and across extended periods of time (e.g., Carpenter et al., 2008, 2009). Consequently, many cognitive and educational psychologists rank testing as among the most potent of all evidence-based learning techniques (e.g., Brown et al., 2014; Roediger & Butler, 2011; Roediger & Karpicke, 2006; Soderstrom & Bjork, 2015). The relative efficacy of the two most prominent types of practice testing, namely *posttesting* and *pretesting*, is the focus of this manuscript.

Posttesting: Retrieval Practice Improves Learning

Most research on test-enhanced learning concerns the use of posttesting, which is commonly known as *retrieval practice*. In investigations of posttesting, the first of which was conducted by Abbott (1909), participants study information (e.g., a text passage, a lecture, a set of facts) and then take a recall test (i.e., posttest) on that information. That posttest might involve multiple-choice, cued recall, free recall, or some other test format. Relative to control conditions wherein a non-testing activity occurs (e.g., a reexposure control such as restudying) or simply no activity occurs at all, posttesting often improves memory as assessed on a subsequent criterial test. A meta-analysis by Rowland (2014; see also Adesope et al., 2017) found that the

typical effect size of that improvement, which is also known as the *testing effect*, is $g = 0.50$, 95% CI [0.42, 0.58]. Testing effects have been successfully demonstrated across a considerable range of materials (for a listing, see Rawson & Dunlosky, 2011), in a variety of authentic educational settings (e.g., McDaniel et al., 2011; Pan, Cooke, et al., 2020), and when posttesting is implemented in a variety of test formats (Rowland, 2014). Further, a meta-analysis by Pan and Rickard (2018) found that the typical size of a transfer effect following posttesting (that is, the ability to apply learning to new contexts, such as to solve application and inference questions, or to recall information on a criterial test that uses a different format) is $d = 0.40$, 95% CI [0.31, 0.50]. Thus, besides enhancing recall, posttesting can improve transfer of learning as well.

Among multiple theoretical accounts of the testing effect that specify underlying mechanisms (for reviews, see Karpicke et al., 2014; Rowland, 2014; van den Broek et al., 2016), a common assumption is that posttesting elicits cognitive processes that non-testing techniques do not. Such processes potentially include elaborative retrieval or mediator generation (Carpenter, 2009; Pyc & Rawson, 2010), wherein semantically related information is encoded along with tested items; contextual feature updating (Karpicke et al., 2014), wherein temporal and episodic features of study and test events are encoded along with tested items; and new memory formation (Rickard & Pan, 2019), wherein a new episodic memory of the test event is formed. By such accounts, posttesting yields qualitatively different memories (e.g., Carpenter, 2009) or separate memories (e.g., Rickard & Pan, 2009) than following non-testing activities (Bjork, 1975), and as a consequence, information becomes more recallable on a criterial test. Importantly, the efficacy of posttesting appears to depend on learners' ability to successfully recall previously learned information; items that are not successfully recalled on a posttest may derive no memorial benefits (Kornell et al., 2011). Information that is not successfully recalled

during posttesting can however be re-acquired if feedback that includes the correct answer is presented (Kornell & Vaughn, 2016).

In a 2007 guide to educational interventions commissioned by the U.S. Department of Education's Institute of Education Sciences (IES), the evidence supporting the efficacy of posttesting was rated as "strong" (Pashler et al., 2007). Similarly, a recent widely-cited review of learning techniques gave posttesting a "high" utility rating (Dunlosky et al., 2013). In both reports, the use of posttests in educational contexts was highly recommended. That advice is partly reflected in everyday practice: Surveys indicate that 62-70% of undergraduate students engage in some form of posttesting (e.g., Hartwig & Dunlosky, 2012; Kornell & Bjork, 2007; Pan et al., 2020a), although more commonly for assessment than to enhance learning.

Pretesting: The Benefits of Errorful Generation

Pretesting, which is also known as *errorful generation* or prequestioning, involves taking practice tests before to-be-learned information is studied, as opposed to afterwards. For example, a student might take a pretest on a textbook chapter before reading it. Owing to a lack of prior knowledge, many errors of commission or omission often occur during such pretests (e.g., in Richland et al., 2009, participants answered less than 10% of pretest items correctly). However, when memory for the correct answers is assessed on a subsequent criterial test, pretesting usually results in better performance than non-testing conditions wherein the correct answers are simply studied from the outset. Thus, pretesting followed by studying can benefit learning. That *pretesting effect* has been demonstrated for science texts (e.g., Richland et al., 2009), video lectures (e.g., Toftness et al., 2017), and foreign language vocabulary (e.g., Potts & Shanks, 2014), as well as at various retention intervals (e.g., Kornell et al., 2009), in laboratory and classroom settings (e.g., Carpenter et al., 2018), and with pretesting in cued recall and multiple-choice format (e.g., Little & Bjork, 2016). Although benefits of pretesting are

commonly limited to memory for directly tested materials (e.g., James & Storm, 2019; Toftness et al., 2017), transfer of learning to materials that were not directly tested has sometimes also been observed (e.g., Carpenter & Toftness, 2017; Pan et al., 2019).

Prominent theoretical accounts of the pretesting effect focus on the generation of errors and/or the subsequent study of correct information (for reviews see Kornell & Vaughn, 2016; Metcalfe, 2017). Accounts that posit mechanisms that occur specifically during pretesting include semantic activation (e.g., Kornell, 2009; Richland et al., 2009), wherein pretest cues activate cue-related knowledge with which targets are encoded; semantic mediation (e.g., Vaughn & Rawson), wherein errors act as mediators between cues and targets; and episodic recollection (e.g., Metcalfe & Huelser, 2020; see also Butterfield & Metcalfe, 2001), wherein the memory of making an error later aids in recollection of the correct answer. Notably, in the case of the semantic mediation and episodic recollection accounts, the generation of errors is critically important. Other accounts focus on processes occurring after pretesting and include increased curiosity (Geller et al., 2017), wherein generating errors prompts learners to search for the correct answers; changes in attention (e.g., McCrudden et al., 2005; McCrudden & Schraw, 2006), wherein pretesting increases learners' focus during subsequent reading or study activities; and other changes in reading or study behaviors (e.g., Bjork et al., 2013; Geller et al., 2017), such as reading with a goal of reducing knowledge gaps. Most of these accounts can be classified as examples of *test-potentiated learning*, wherein testing improves the effectiveness of subsequent reading or study activities (Arnold & McDermott, 2013; Chan et al., 2018). Unlike posttesting, the benefits of pretesting do not rely on retrieval success during practice testing; as previously noted, a low rate of retrieval success is common. Rather, in order for a pretesting effect to manifest, it is crucial that there be an opportunity to learn the correct answers after retrieval attempts are made, whereas with posttesting, no such opportunities are necessary except

in cases of unsuccessful retrieval.

The first investigations of pretesting were conducted over four decades ago (e.g., Rickards, 1976), but most well-controlled studies on the subject were only published in the last decade. Hence, many learning scientists do not currently rate pretesting as comparable to posttesting in applicability or effectiveness. For instance, the aforementioned IES guide (Pashler et al., 2007) rated the level of evidence regarding pretesting as "low," although it noted several studies showing promising results. Pretesting is entirely unmentioned in Dunlosky et al.'s (2013) review of learning techniques. Further, although some researchers recommend using pretesting in educational contexts (e.g., Bjork & Bjork, 2014), a recent large-scale survey found that 87% of undergraduate students do not use practice questions to do so (Pan et al., 2020a). Thus, both the acceptance and prevalence of pretesting substantially lags behind that of posttesting.

Is Pretesting Competitive with Posttesting?

The foregoing research suggests that an instructor or student intending to foster test-enhanced learning, such as to improve the learning of a text passage or a book chapter, might profitably use pretesting or posttesting. Given those options, a practical question arises: which is more beneficial? In the current literature, the relative efficacy of pretesting and posttesting has yet to be established. In fact, both types of practice testing have seldom been directly compared under controlled circumstances, and the relevant studies to date have featured certain design factors—including the intermixing of test and study activities, no controls for reading time, prior study of to-be-learned materials, and the use of immediate correct answer feedback—that complicate interpretation.

The first studies to compare pretesting and posttesting used embedded *adjunct questions* wherein test questions were presented throughout a text passage or chapter (for reviews see

Anderson & Biddle, 1975; Hamaker, 1986). For example, Rothkopf and Bibiscos (1967) embedded two questions before or after every three pages of a two-chapter biology text and had participants alternate between reading and answering questions before taking a criterial test. The questions referring to content yet to be read constituted pretesting, whereas the questions referring to content that had already been read constituted posttesting. Studies using this methodology (e.g., Frase, 1967, 1968; Rickards, 1976; Swenson & Kulhavy, 1968; see also Rothkopf, 1966; Sagaria & Di Vesta, 1978) reported mixed results, although a posttesting advantage was repeatedly observed (cf. Rickards, 1976,1977). However, the intermixing of test questions and text that occurred in those studies—that is, testing that is *interpolated* with studying—is now known to alter the efficacy of those test questions and attendant cognitive processes (e.g., reductions in proactive interference) relative to cases wherein testing is not repeatedly mixed with studying (e.g., Davis et al., 2017; Pan et al., 2020b; Szpunar et al., 2008; Wissman et al., 2011). Hence, although studies with embedded adjunct questions provide insights into the consequences of mixing testing and studying, the effects of non-interpolated pretesting or posttesting are not addressed. Additionally, because many studies of adjunct questions do not control for time-on-task, differences in reading time between the pretesting and posttesting conditions may have also been a factor.

More recently, McDaniel et al. (2011, Experiments 2a and 2b) had 8th grade students take quizzes before or after classroom science lessons. Notably, interpolated testing was not used. Post-lesson but not pre-lesson quizzing enhanced performance on subsequent unit and final exams, which was interpreted as evidence that pretesting is less effective than posttesting. However, the pre-lesson quizzes occurred *after* assigned readings had already been completed (following reading, the average performance on those quizzes was relatively high, at over 50%), and as such those quizzes arguably constituted posttesting rather than pretesting. Further,

immediate correct answer feedback was provided during quizzing. By some accounts, the addition of feedback to pretesting may reduce participants' attention to subsequently presented materials given that the answers are already known (Sana et al., in press), and in the case of incorrect responding, may reduce motivation to learn (Latimier et al., 2019). Alternatively, feedback may enhance learning by serving as an extra study opportunity (e.g., Hausman & Rhodes, 2018). Thus, although the results of McDaniel et al. suggest that pretesting may be ineffective in an authentic educational context, it is not clear whether the same results would be obtained if learners had minimal preexisting knowledge prior to pretesting and did not receive feedback. Finally, in another study, Latimier et al. (2019) had participants take interpolated pretests or posttests with feedback during an online biology lesson and found that posttesting outperformed pretesting on a 7-day delayed criterial test. However, two of the same concerns as in aforementioned studies, namely the use of interpolated testing and immediate correct answer feedback, apply to that study as well.

Overall, although the evidence to date suggests that posttesting may have an edge when it comes to promoting learning, the question of whether the two test types are competitive with one another remains to be examined under circumstances that reflect the most common instantiations of both (and for pretesting especially), including without the intermixing of test and study trials, in the absence of immediate correct answer feedback, and when accompanied by a single opportunity to study to-be-learned materials. Under such circumstances, different results might be obtained. Theoretically, when implemented in isolation (e.g., without feedback), the effectiveness of posttesting may be constrained by the ability to successfully retrieve previously studied information (Rowland, 2014; Smith & Karpicke, 2014), which is itself constrained by the degree with which that the information was initially encoded, whereas with pretesting, the encoding of information after a pretest may be enhanced via test-potentiated learning (and

retrieval success is not important). If so, then pretesting may have an advantage in cases wherein the degree with which target information is encoded is a critical factor.

The Present Study

Our investigation of the efficacy of pretesting versus posttesting focused on the following simplest case scenario: a single practice test that is taken before or after reading an expository text passage. That scenario facilitated an arguably purer comparison of the two test types than in prior research (that is, focused on isolated implementations of pretesting and posttesting and avoiding or addressing the aforementioned factors that may influence the relative effectiveness of each). Crucially, unlike prior studies, we did not use interpolated testing, and in most cases, did not provide immediate correct answer feedback. The text passages were also unfamiliar to our participants, thus reducing effects of outside knowledge and ensuring a high rate of guessing during pretesting. Both passages had been used previously in other studies to demonstrate robust pretesting or posttesting effects, including relative to a no-testing and restudy control (e.g., Little, 2011; Little & Bjork, 2016; Little et al., 2012).

Within each of five experiments, participants read one or two text passages, each accompanied by a single pretest or posttest. Afterwards, they took a criterial test during which retention and transfer of learning was assessed. With an aim of investigating the generality of any differences between pretesting and posttesting, we first investigated testing in multiple-choice format (Experiment 1), then addressed testing in cued recall format (Experiment 2), testing with and without immediate correct answer feedback (Experiment 3), and testing across a 48-hour retention interval (Experiment 4). In a final experiment, we investigated the theoretical role of test-potentiated learning in driving differences between pretesting and posttesting. That experiment involved conditions wherein participants completed a single practice test and two readings of a text passage (Experiment 5).

Experiment 1

The first experiment investigated the relative benefits of taking a multiple-choice pretest or posttest on the learning of expository text passages.

Method

Participants

Participants were recruited from Amazon Mechanical Turk (MTurk) and compensated with USD $5 each. Participation was limited to MTurk workers from North America that were fluent in English and had an approval rate of 95% or higher on prior MTurk studies. A power analysis using the G*Power program (Faul et al., 2007) indicated that a sample of 36 participants would be needed for 95% power to detect a medium-sized main effect of test-enhanced learning (f = 0.25) in a 2x2 within-subjects design at ± = 0.05 (based on a posttesting effect size of g = 0.50, per Rowland, 2014). To address concerns about sufficient statistical power (including with respect to interactions; e.g., Gelman, 2018; Simonsohn, 2014) and to account for expected attrition due to potential technical and other issues, we set a substantially larger sample size target for this experiment, 150, and posted slots in excess of that amount to ensure that it was reached. One hundred and seventy-four participants (M_{age} = 35.2 years, 58% male) completed the experiment without technical problems and were included in the analyses (in all experiments, evidence of technical issues or unexpected distractions, as indicated by server logs and/or responses to debriefing questions, were used to identify participants that did not complete the experiment as instructed). All experiments in this study were approved by the Institutional Research Ethics Boards (IRB) of Athabasca University or the University of California, Los Angeles, and all participants gave informed consent before participating.

Design

We used a 2 (Test Type: Pretest vs. Posttest) x 2 (Question Type on the criterial test:

Tested vs. Untested) within-subjects design. Each participant completed two experimental blocks. Within each block, they read an expository text passage, took a pretest or a posttest on that passage, and then completed a 5-minute delayed criterial test. Assignment of passage to test type (pretest or posttest), passage order, test order (pretested passage first or posttested passage first), and question set per passage (set A or set B) were counterbalanced.

Materials

The stimuli were two 1,100-word encyclopedia-style expository text passages originally developed by Little (2011) and later adapted for use in Little et al. (2012) and Little and Bjork (2016). The passages had a Flesch-Kincaid grade level of 10.5 and described Yellowstone National Park and the planet Saturn, respectively. Both passages were constructed around ten categories of factual information (e.g., for the Saturn passage: moons, probe visits, other planets, etc.) and featured at least four exemplars per category, all of which were presented in the passage and commonly in close to temporal proximity to one another. There was a pair of multiple-choice questions for each category, with each question per pair drawing on facts taken almost verbatim from the passage and having the same four answer options (i.e., exemplars) but different correct answers. For example, the two questions about Saturn's moons were, "*What is Saturn's largest moon?"* and "*What is Saturn's second largest moon?*", with both questions having the same answer options (*Titan, Rhea, Mimas,* and *Enceladus*) but different correct answers (e.g., *Titan* and *Rhea* for the former and latter questions, respectively). The answer options for each question were competitive with one another and required careful consideration in order to determine the correct answer, which is a feature that is known to enhance the potency of such test questions (Little & Bjork, 2016). Text passage excerpts and example questions are presented in Appendices A and B, respectively. All materials are accessible at the Open Science Framework (https://osf.io/zd86x/).

The 20 questions per passage were divided into two 10-question sets, with one question per category randomly assigned to each set (forming sets A and B; see Appendix B for examples of the sets from both text passages). For each participant and for each passage, eight questions from one set were used during the initial learning phase (for pretesting or posttesting) and all questions from both sets were used on the criterial test (for a total of eight *Tested* questions, eight *Untested* questions, and four *Control* questions). Tested questions were identical to those used during practice testing and assessed memory for previously tested content. Untested questions were not used on the practice test but drew from the same categories of information and assessed a form of transfer of learning (i.e., to previously read but not tested materials; Pan & Rickard, 2018). In other words, Untested questions addressed whether taking a test on one exemplar from a category could enhance memory for other exemplars of that category. Such transfer could be considered to be relatively "near" on a near-versus-far transfer spectrum (Barnett & Ceci, 2002) given that the categories targeted by Untested and Tested questions were identical (e.g., whereas a Tested question addressed Saturn's largest moon, an Untested question addressed Saturn's second largest moon).

Control questions drew from different categories as the other questions and were not used for pretesting or posttesting. These questions were primarily included to address potential effects of a differential lag-to-test between the reading of the passages and the criterial test (that is, after the reading of a passage, precisely 5 minutes elapsed in the case of the pretesting condition, whereas more than that time elapsed in the posttesting condition given the intervening posttest). Control questions were analyzed separately for each experiment and further served as a non-testing reference condition for supplementary meta-analyses that are detailed later in this manuscript. Between the Tested, Untested, and Control questions, all ten categories of information addressed in each passage (which, with the exception of the opening sentences,

comprised the vast majority of each passage) were tested, and between the Tested and Untested questions, two exemplars of each category were tested (on the practice and/or criterial tests).

Procedure

All experiments were conducted using LimeSurvey (Limesurvey GmbH) and accessed via internet browser. Participants first read instructions stating that they were to read two text passages about different topics and that they were to answer practice questions immediately before or after reading each passage. They were asked to guess if they did not know the correct answer to any question and to expect a subsequent test on their knowledge of each passage.

As depicted in Figure 1, the experiment consisted of two blocks, with each block devoted to one text passage and having three phases: the learning phase, a distractor task, and the criterial test. During the learning phase, participants read a passage at their own pace and responded to eight randomly ordered multiple-choice test questions about that passage. These questions were presented before (pretesting) or after (posttesting) the passage. The answer options for each question were randomized on each trial. Afterwards, they completed a 5-minute distractor task (a backwards digit span task) and then the criterial test. All test questions were self-paced and presented one at a time in random order. As with practice testing, the answer options for each criterial test question were randomized on each trial. No feedback was provided. A second block immediately followed the first and featured the same procedures excepting a change in the type of practice test (i.e., if a participant took a pretest during the first block, then that participant would take a posttest in the second block, or vice versa) and a different text passage. After the second block, participants were debriefed, after which the experiment concluded.

Results and Discussion

Descriptive statistics for learning phase and criterial test performance for Experiments 1-4 are reported in Tables 1 and 2, respectively. In this and subsequent experiments, analyses were

performed on data combined across text passages and any other counterbalancing factors.

Learning Phase

Practice test performance. As expected, performance on the posttests (M = 61%, SD = 24%) was greater than on the pretests (M = 33%, SD = 18%). When considering that the expected accuracy rate for random guessing on the pretests (given the use of the multiple-choice format) is 25%, it is evident that most participants made many erroneous responses during pretesting. That pattern confirms low prior knowledge of passage content and is consistent with the majority of the pretesting literature (cf. Latimier et al., 2019; McDaniel et al., 2011). Further, the significantly higher performance on the posttests implies that substantial learning occurred during the reading of the passages.

Reading time. The amount of time that participants spent reading the text passage in the pretest (M = 5.0 minutes, SD = 2.8 minutes) and posttest conditions (M = 5.3 minutes, SD = 3.3 minutes) was not significantly different, $t(173) = 1.50$, $p = .13$.

Criterial Test

Control questions. Performance on control questions was similar regardless of whether participants engaged in pretesting (M = 56%, SD = 29%) or posttesting (M = 55%, SD = 30%), $t(173) = .11$, $p = .92$. This result suggests that any effects of the differential lag-to-test between conditions were minor.

Tested and Untested questions. Results (means and standard deviations for the pretest and posttest conditions) are summarized in Table 2. A repeated measures Analysis of Variance (ANOVA) on criterial test scores was conducted to compare the effect of Test Type (Pretest vs. Posttest) on the learning of directly tested information and untested information drawn from the same content categories. There was a significant main effect of Test Type, indicating that performance on the criterial test was better after pretesting than posttesting, $F(1, 173) = 12.33$, p

= .001, $\cdot_p^2$ = 0.07. There was also a significant main effect of Question Type, indicating that performance on the criterial test was better for Tested than Untested questions, $F(1, 173) = 18.26$, $p < .001$, $\cdot_p^2 = 0.10$. The latter result indicates that performance was better on questions that were previously attempted versus entirely novel, as should be expected given prior exposure to those questions during the learning phase (and is consistent with the conclusion that test-enhanced learning is strongest for previously tested content, as noted in Pan & Rickard, 2018). Further, there was also a non-significant interaction between Test Type and Question Type, $F(1,173) = .66$, $p = .42$, which indicates that the advantage of pretesting over posttesting was maintained across both types of materials, although it was numerically smaller for Untested questions. The advantage of pretesting over posttesting (percent increase) was $M = 11\%$ and $M = 7\%$ for Tested and Untested questions, respectively.

In a supplementary analysis we investigated whether the presentation order of pretested versus posttested text passages, which was counterbalanced across participants, had any effect on criterial test performance. That analysis involved an ANOVA similar to the one discussed above but with an added factor of Test Order (Pretested first vs. Posttested first). The main effect of Test Order was significant, indicating that criterial test performance was better when posttesting occurred first ($M = 65\%$, $SD = 24\%$) compared to when pretesting occurred first ($M = 60\%$, $SD = 24\%$), $F(1,172) = 3.95$, $p = .049$, $\cdot_p^2 = 0.02$. However, the main effect of Test Type was still observed, indicating that performance on the criterial test was better after pretesting than posttesting, $F(1,172) = 12.53$, $p = .001$, $\cdot_p^2 = 0.07$. There were no significant interactions of Test Order and Test Type or Question Type (*p*s e .05). Thus, the patterns observed in this experiment were not attributable to the presentation order of the type of test.

Overall, the results of Experiment 1 indicate that pretesting can promote learning to a greater extent than posttesting, at least for learning expository text passages, with multiple-

choice practice and criterial tests, and when the criterial test occurs after a 5-minute retention interval.

Experiment 2

In the second experiment we investigated potential effects of test format—that is, cued recall as opposed to multiple-choice—on the relative benefits of pretesting versus posttesting. In the retrieval practice literature, both multiple-choice and cued recall formats, which are among the most common formats used, can yield substantial testing effects (Rowland, 2014), although some direct comparisons of those formats have found evidence of differing effectiveness (e.g., Kang, et al., 2007; Little et al., 2012; Smith & Karpicke, 2014). Such patterns, however, have not been consistent across studies. In the pretesting literature, both multiple-choice and cued recall formats can yield substantial pretesting effects (e.g., Richland et al., 2009; St. Hilaire & Carpenter, 2020), but relatively few studies have directly compared formats. In one such comparison, Little and Bjork (2016) found that the multiple-choice format was more efficacious than cued recall when the multiple-choice answer options were competitive with one another, with those answer options presumably promoting greater processing of to-be-learned content.

Method

Participants

Participants were recruited from MTurk using the same criteria and with the same compensation as in Experiment 1. A power analysis using the G*Power program indicated that a sample of approximately 76 participants would be needed to detect a medium-sized effect (f = 0.25) using a 2x2x2 mixed design with ± = 0.05 and 95% power. We again recruited in excess of that amount, releasing slots for up to 320 participants. Two hundred and seventy-three participants (M_{age} = 36.8 years, 53% male) completed the entire experiment without problems and were included in the analyses.

Design

We used a 2 (Test Type: Pretest vs. Posttest) x 2 (Test Format: Multiple-choice vs. Cued recall) x 2 (Question Type on the criterial test: Tested vs. Untested) mixed design wherein Test Type and Test Format were analyzed as between-subjects factors and Question Type was analyzed as a within-subjects factor. As in the preceding experiment, all participants read two passages, one preceded by a pretest and the other followed by a posttest. For each participant, one practice test involved cued recall (i.e., short answer) and the other test involved multiple-choice. Hence, Test Type and Test Format were treated as between-subjects factors in the analysis. The order of the passages, tests, and formats, as well as the question set used for each passage, were all counterbalanced. The criterial test remained in multiple-choice format.

Materials and Procedure

All materials and procedures were identical to Experiment 1 with one exception: For the cued recall tests, participants did not choose among a series of answer alternatives; rather, they were instructed to type their answer into a textbox that appeared below the question. The criterial test remained entirely multiple-choice.

Scoring

In this and subsequent experiments, all cued recall data were computer scored using a similarity-matching method wherein responses were analyzed in Microsoft Excel using the Fuzzy Lookup (Microsoft Research, Redmond, WA) add-in (Pan & Rickard, 2017). That add-in compared participants' responses to a master list of correct answers, with those responses having to be a very close match to the spelling of the correct answers to be counted as correct.

Results and Discussion

Learning Phase

Practice test performance. As in the preceding experiment, performance on the posttests

(*M* = 54%, *SD* = 23%) was greater than on the pretests (*M* = 22%, *SD* = 18%). Further, that pattern was apparent regardless of whether multiple-choice or cued recall format was used (for format-specific results, see Table 1).

Reading time. The amount of time that participants spent reading the text passage in the pretest and posttest conditions (*M* = 6.0 minutes, *SD* = 4.0 minutes across all conditions) was not significantly different in the case of multiple-choice testing, $t(271) = 1.02$, $p = .31$, or cued recall testing, $t(258) = 1.47$, $p = .14$.

Criterial Test

Control questions. Performance on control questions following pretesting (*M* = 57%, *SD* = 27%) and posttesting (*M* = 55%, *SD* = 30%) was similar, $t(544) = .49$, $p = .63$, just as in Experiment 1. Thus, any effects of differential lag-to-test appear to have been minor.

Tested and Untested questions. Table 2 summarizes the descriptive statistics for each condition by question type and test type. A 2 (Test Type: Pretest vs. Posttest) x 2 (Test Format: Multiple-choice vs. Cued recall) x 2 (Question Type: Tested vs. Untested) mixed-factors ANOVA on criterial test scores yielded a significant main effect of Test Type, $F(1, 542) = 4.71$, $p = .030$, $\cdot_p^2 = 0.01$, indicating that performance was greater following pretesting than posttesting. That result replicates the pretesting advantage that was observed in Experiment 1. There was also a significant main effect of Question Type, $F(1, 542) = 82.89$, $MSE = .03$, $p < .001$, $\cdot_p^2 = 0.13$, which indicates that performance was better on previously attempted versus novel questions, as was also observed in Experiment 1. The main effect of Test Format was non-significant, $F(1,542) = .53$, $p = .47$, which indicates that criterial test performance was unaffected by the use of practice tests in multiple-choice as opposed to cued recall format. That finding contrasts with the conclusions of Little et al. (2012) and Little and Bjork (2016), in which multiple-choice tests were found to be more effective for posttesting and pretesting, respectively.

All interactions were non-significant ($p > .05$). Across conditions, the advantage of pretesting over posttesting (percent increase) was $M = 8\%$ and $M = 5\%$ for Tested and Untested questions, respectively.

Further, a supplementary analysis of Test Order analogous to that conducted for the preceding experiment revealed no significant main effect of Test Order, $F(1, 538) = .37$, $p = .54$, and no significant interactions between Test Order and Test Type, Test Format, or Question Type (*p*s e .05). Thus, the patterns observed in this experiment were not attributable to the presentation order of the test types.

Overall, the results of Experiment 2 replicate and extend those of Experiment 1: Pretesting in either multiple-choice or cued recall format yielded greater learning than posttesting. It thus appears that test format, at least with respect to the two common formats investigated here, is not determinative of the relative effectiveness of pretesting versus posttesting.

Experiment 3

In the first two experiments, pretesting and posttesting were implemented without immediate *correct answer feedback*—that is, the presentation of answers once participants have entered their responses. In Experiment 3, we investigated whether adding such feedback could influence the relative benefits of pretesting versus posttesting. In the retrieval practice literature, feedback can increase testing effect magnitude (Rowland, 2014), ameliorate the negative effects of recalling incorrect information (e.g., Butler & Roediger, 2008), compensate for low retrieval success (e.g., Kang et al., 2007), and improve transfer of learning (e.g., Pan et al., 2018). Thus, we expected that feedback would enhance the benefits of posttesting.

In the pretesting literature, although the opportunity to study correct answers is required for a pretesting effect to manifest, whether that opportunity takes the form of immediate correct

answer feedback has varied according to the materials being learned (for discussion see Kornell, 2014). In the case of text passages, participants typically read the relevant text passage at some point after pretesting (e.g., Richland et al., 2009) and no formal feedback is provided, whereas for more semantically impoverished materials such as paired associates, immediate correct answer feedback is necessary (e.g., Hays et al., 2013; Vaughn & Rawson, 2012). In at least one pretesting study with text passages, however, the provision of correct answer feedback was manipulated (Hausman & Rhodes, 2018), and that feedback enhanced the magnitude of the pretesting effect. We hypothesized that adding feedback in Experiment 3, albeit relatively uncommon for pretesting with text passages, might enhance learning for the case of Tested questions. As previously noted, however, two alternate possibilities are that feedback might impair learning by reducing the need to search for the correct answers when the passage is subsequently read (Sana et al., in press; St. Hilaire & Carpenter, 2020) or reduce motivation to learn (Latimier et al., 2019).

Method

Participants

Participants were recruited from MTurk in the same manner as in the prior experiments and compensated with USD $5 each. A power analysis using the G*Power program indicated that approximately 90 participants would be needed to detect a medium-sized effect (f = 0.25) using a 2x2x2x2 mixed design with ± = 0.05 and 95% power. We again aimed for a sample size beyond that amount, opening slots for up to 530 participants. Four hundred and ninety-two participants (M_{age} = 38.8 years, 53% female) completed the experiment and were included in the analyses.

Design

We used a 2 (Test Type: Pretest vs. Posttest) x 2 (Test Format: Multiple-choice vs. Cued

recall) x 2 (Feedback: yes or no) x 2 (Question Type on the criterial test: Tested vs. Untested) mixed design with all factors manipulated between-subjects excepting Question Type, which was manipulated within-subjects.

Materials and Procedure

Only the Yellowstone National Park passage and its two question sets were used in this experiment (i.e., each participant read and took a practice test on one text passage). In the learning phase, all participants read that passage paired with an eight-question pretest or posttest in multiple-choice or cued recall format. Participants that received feedback did so immediately in the form of the correct answer being presented on a subsequent screen after they selected or entered a response to a given practice question. As with the practice test questions, feedback was self-paced. Participants then completed a 5-minute distractor task and a 20-question, self-paced criterial test that was patterned after that used in the prior experiments.

Results and Discussion

Learning Phase

Practice test performance. Performance on the posttests (*M* = 57%, *SD* = 24%) was greater than on the pretests (*M* = 27%, *SD* = 19%). Further, that pattern was apparent regardless of whether multiple-choice or cued recall format was used, or whether feedback was provided or not (for format- and feedback-specific results, see Table 1). These patterns were similar to those observed in the preceding experiments.

Reading time. The amount of time spent reading the text passage (*M* = 5.9 minutes, *SD* = 5.0 minutes across all conditions) did not significantly differ between pretesting and posttesting conditions for each format-feedback combination (*t*s d 0.64, *p*s e .52).

Criterial Test

Control questions. Performance on control questions was similar between the pretest (*M*

= 50%, *SD* = 28%) and posttest (*M* = 54%, *SD* = 28%) conditions, $t(490) = 1.58$, $p = .11$, matching the patterns observed in the prior experiments.

Tested and Untested questions. A 2 (Test Type: Pretest vs. Posttest) x 2 (Test Format: Multiple-choice vs. Cued recall) x 2 (Feedback: yes or no) x 2 (Question Type: Tested vs. Untested) mixed-factors ANOVA on criterial test scores (see Table 2) yielded a significant main effect of Test Type, $F(1, 484) = 5.16$, $p = .024$, $\cdot_p^2 = 0.01$, indicating that performance on the criterial test was better with pretests than with posttests, and a significant main effect of Question Type, $F(1, 484) = 464.49$, $p < .001$, $\cdot_p^2 = 0.49$, indicating that performance was better on Tested than Untested questions. The main effect of Test Format was not significant, $F(1, 484) = .55$, $p = .46$. Overall, these patterns mirror those from the preceding experiments: Pretesting was more effective than posttesting, criterial test performance was better for previously seen than for new questions, and the test format that was used during practice testing did not influence the results. Across conditions, the advantage of pretesting over posttesting (percent increase) was *M* = 5% and *M* = 6% for Tested and Untested questions, respectively.

Importantly, there was also a significant main effect of Feedback, $F(1, 484) = 18.06$, $p < .001$, $\cdot_p^2 = 0.04$, indicating that criterial test performance was better when correct answer feedback was provided during practice testing. That main effect was qualified by a significant Feedback by Question Type interaction, $F(1, 484) = 103.66$, $p < .001$, $\cdot_p^2 = 0.18$. Follow-up tests revealed that performance on Tested questions was greater when practice tests included feedback as opposed to no feedback, $t(490) = 10.89$, $p < .001$, $d = 0.99$, and performance on Untested questions was similar regardless of whether practice tests included feedback or not, $t(490) = 1.67$, $p = .097$. Further, performance was better on Tested than Untested questions when feedback was provided, $t(254) = 22.67$, $p < .001$, $d = 1.38$, than when it was not, $t(236) = 8.01$, $p < .001$, $d = 0.49$. Overall, these results indicate that correct answer feedback was more

beneficial for Tested than Untested questions, which is plausible given that the information presented as feedback was directly relevant for Tested questions only.

The lack of any significant interactions involving Feedback with Test Type, $F(1, 484) = 1.27$, $p = .26$, and Feedback with Test Format, $F(1, 484) = .00$, $p = .99$, suggests that the provision of feedback did not affect pretesting versus posttesting, or multiple-choice versus cued recall testing, in different ways. Rather, feedback improved performance on Tested questions across all test formats in this experiment. Therefore, we did not find evidence of a moderating role of feedback on the advantage of pretesting over posttesting, nor did we observe indications that feedback reduced the efficacy of pretesting.

Experiment 4

The foregoing experiments all indicate that pretesting can be more effective than posttesting at promoting the learning of expository texts. However, a limitation of those experiments is that the criterial test occurred after a period of only 5 minutes. Although benefits of posttesting relative to non-testing conditions have been observed at similarly short retention intervals (e.g., Rowland & DeLosh, 2015), those benefits can be more apparent at intervals greater than 24 hours (e.g., Rowland, 2014 reported an effect size of $g = 0.41$, 95% CI [0.31, 0.51] when the retention interval was less than one day and $g = 0.69$, 95% CI [0.56, 0.81] when it was longer than one day). In the pretesting literature, learning is commonly assessed after a retention interval of just a few minutes, but the pretesting effect also manifests at intervals ranging from 24 hours (e.g., Kornell et al., 2009) to one week (e.g., Richland et al., 2009; see also Pan et al., 2019). Moreover, from an educational standpoint, learning ideally should persist over an extended period of time. Accordingly, Experiment 4 was designed to assess whether the advantage of pretesting over posttesting survives over a 48-hour retention interval, and whether the other factors investigated in the prior experiments, namely test format and feedback,

influence that advantage.

Method

Participants

For logistical reasons, we recruited undergraduate students for this experiment from the participant pool at a large public university in North America in exchange for course credit. Given similarities in experimental design, the same power analysis from Experiment 3 applied. Two-hundred and thirty-six undergraduate students (M_{age} = 20.5 years, 73% female) completed the entire experiment and were included in the analyses (data was only analyzed for participants that returned to complete the second session).

Design, Materials, and Procedure

The design, materials, and procedure were identical to that of Experiment 3 except for two differences: The learning phase and criterial test phases were separated by 48 hours and all participants completed both parts of the experiment in a supervised lab setting. During the first session, the experimental block was identical to that of the preceding experiment except for the absence of the criterial test, which was withheld until the second session.

Results and Discussion

Learning Phase

Practice test performance. Performance on the posttests (M = 41%, SD = 21%) was greater than on the pretests (M = 17%, SD = 10%). As in Experiment 3, that pattern was apparent regardless of whether multiple-choice or cued recall format was used, or whether feedback was provided or not (for format- and feedback-specific results, see Table 1). Overall, these patterns were again similar to those observed in the preceding experiments, albeit notably lower than the corresponding results in Experiment 3.

Reading time. The amount of time spent reading the text passage (M = 6.2 minutes, SD

= 2.8 minutes across all conditions) did not significantly differ between pretesting and posttesting conditions for most format-feedback combinations (*t*s d 1.17, *p*s e .25). The sole exception was the case of cued recall with feedback, for which the average reading time in the pretesting condition (M = 4.6 minutes, SD = 2.1 minutes) was significantly less than that for the posttesting condition (M = 6.1 minutes, SD = 2.7 minutes), $t(53) = 2.39$, $p = .020$, $d = 0.62$. That finding is consistent with the possibility that adding immediate correct answer feedback to pretesting reduces participants' need to search for answers during the subsequent reading of a text passage, and hence shortens the time spent reading that passage. It should however be reiterated that no significant reading time difference was observed among the cued recall with feedback conditions in Experiment 3.

Criterial Test

Control questions. Performance on control questions was similar following pretesting (M = 36%, SD = 27%) and posttesting (M = 34%, SD = 27%), $t(234) = .67$, $p = .50$, matching the patterns observed in the prior experiments.

Tested and Untested questions. We conducted an ANOVA on criterial test scores (see Table 2) that was identical to that performed for Experiment 3. Crucially, there was a significant main effect of Test Type, $F(1, 228) = 9.82$, $p = .002$, $\cdot_p^2 = 0.04$, indicating that pretesting was more effective than posttesting. There were also significant main effects of Question Type, $F(1, 228) = 336.63$, $p < .001$, $\cdot_p^2 = 0.60$, and Feedback, $F(1, 228) = 13.50$ $p < .001$, $\cdot_p^2 = 0.06$; these indicated that performance was better on Tested than Untested questions and when feedback was provided than when it was not, respectively. The main effect of Test Format was not significant, $F(1, 228) = 3.06$, $p = .081$. All of these patterns mirrored those observed in the preceding experiments. Across conditions, the advantage of pretesting over posttesting (percent increase) was M = 18% and M = 1% for Tested and Untested questions, respectively.

There were also four significant two-way interactions. A significant Test Type by Question Type interaction, $F(1, 228) = 13.88$, $p < .001$, $\cdot_p^2 = 0.06$, reflected an advantage for pretesting over posttesting for Tested questions, $t(234) = 4.34$, $p < .001$, $d = 0.57$, but not for Untested questions, $t(234) = .06$, $p = .95$. Thus, the pretesting advantage after an extended retention interval in Experiment 4 was limited to previously seen questions. A significant Question Type by Test Format interaction, $F(1, 228) = 5.97$, $p = .015$, $\cdot_p^2 = 0.03$, reflected the fact that multiple-choice and cued recall tests yielded comparable performance for Tested questions, $t(234) = .39$, $p = .70$, but not for Untested questions (wherein there was an advantage for multiple-choice testing, $t(234) = 2.87$, $p = .005$, $d = 0.40$). That result mirrors Little et al.'s (2012) and Little and Bjork's (2016) findings, although we did not observe an advantage for either test format in Experiments 2 and 3 with a 5-minute retention interval.

The remaining two-way interactions involved the provision of feedback. There was a significant Feedback by Question Type interaction, $F(1, 228) = 65.90$, $p < .001$, $\cdot_p^2 = 0.22$, which reflected feedback boosting performance for Tested questions, $t(234) = 7.73$, $p < .001$, $d = 1.11$, and impairing performance for Untested questions, $t(234) = 2.47$, $p = .014$, $d = 0.30$. A similar pattern (but absent any impairment) was observed in Experiment 3. As in that experiment, these results can be partly ascribed to the greater relevance of feedback for Tested than Untested questions. There was also a significant Feedback by Test Format interaction, $F(1, 228) = 4.16$, $p = .043$, $\cdot_p^2 = 0.02$, which reflected an advantage for multiple-choice testing over cued recall testing when no feedback was provided, $t(116) = 2.83$, $p = .006$, $d = 0.55$, and no advantage when it was provided, $t(116) = .24$, $p = .81$. That result is consistent with the possibility that feedback can equalize the efficacy of cued recall and multiple-choice tests, as has been shown for posttesting (e.g., Kang et al., 2007).

Finally, the results of Experiment 4 were qualified by a significant three-way interaction

between Test Type, Question Type, and Feedback, $F(1, 228) = 5.13$, $p = .024$, $\cdot_p^2 = 0.02$. Follow-up tests revealed patterns consistent with the foregoing analyses, including pretesting being more effective than posttesting for Tested questions (*p*s d .027) and being equally effective for Untested questions (*p*s e .49) irrespective of feedback, as well as performance being better on Tested than Untested questions following pretesting (*p*s d .001) and posttesting (*p*s d .048), irrespective of feedback. These results reinforce the conclusion that the advantage of pretesting over posttesting at a 48-hour retention interval is limited to previously seen questions. Moreover, adding feedback to pretesting enhanced performance for Tested questions, $t(122) = 4.89$, $p < .001$, $d = 0.88$, but not for Untested questions, $t(22) = 1.13$, $p = .27$, whereas adding feedback to posttesting enhanced performance for Tested questions, $t(110) = 6.73$, $p < .001$, $d = 1.26$, and worsened performance for Untested questions, $t(110) = -2.40$, $p = .018$, $d = 0.45$. These patterns are also consistent with the greater relevance of feedback for Tested versus Untested questions.

Overall, the results of Experiment 4 affirm the robustness of the advantage of pretesting over posttesting, which was observed after a 48-hour retention interval. However, that advantage was limited to previously seen questions and did not extend to novel questions. Further, the provision of feedback improved performance on Tested but not Untested questions, just as in Experiment 3, but in this case it yielded deleterious effects for Untested questions in the posttesting condition. That pattern is further discussed in the General Discussion.

Experiment 5

The results of Experiments 1-4 provide substantial evidence that pretesting can be highly competitive, and in a range of circumstances even more potent, than posttesting. For the final experiment, we investigated a potential theoretical explanation of the observed pretesting advantage: Taking a pretest may alter the encoding of the text passage that follows it, resulting in

test-potentiated learning, whereas with posttesting, there is no opportunity for such learning to take place (given that participants no longer have access to the text passage). That test-potentiated learning advantages the pretesting condition relative to the posttesting condition.

To test this account, we used a design wherein participants read the same text passage twice during the learning phase. That reading occurred in three different between-subjects conditions: (a) twice after a pretest, (b) twice before a posttest, or (c) once before *and* once after a test. The first two conditions were largely identical to the pretesting and posttesting conditions of the prior experiments, respectively, except that there was a second reading of the text passage immediately after the first reading. In the third condition, which we labeled the *Read-Test-Read* condition, the first reading of the passage and the test that occurred immediately afterwards resembled the posttesting procedure used in the prior experiments (and, more broadly, conventional approaches to posttesting), whereas the second reading of the passage, which presumably was informed by the recent experience of taking a practice test (which, at the moment it was taken, could be described as a posttest), was an opportunity for test-potentiated learning to occur. If so, then criterial test performance in the Read-Test-Read condition should be more competitive with pretesting than the (conventional) posttesting condition.

Method

Experiment 5 was preregistered at: https://aspredicted.org/blind.php?x=js6kb7.

Participants

Similar to the prior experiment, participants were undergraduate students recruited from the subject pool at a large public university in North America and compensated with course credit, but in this case participation occurred online. A power analysis using the G*Power program (Faul et al., 2007) indicated that a sample of 189 participants would be needed for 95% power to detect a medium-size main effect (f = 0.25) in a 3x2 between-subjects design at ± =

0.05. We again recruited well in excess of that amount. Three-hundred and ninety-eight participants (M_{age} = 19.9 years, 76% female) completed the entire experiment without any technical problems and were included in the analyses.

Design, Materials and Procedure

We used a 2 (Test Type: Pretest vs. Posttest vs. Read-Test-Read) x 2 (Question Type on the criterial test: Tested vs. Untested) mixed design wherein Test Type was manipulated between-subjects and Question Type was manipulated within-subjects. Each participant was randomly assigned to one test type. Similar to Experiment 1, only the multiple-choice test format was used and no feedback was provided throughout the experiment. The materials were the same Yellowstone National Park passage as in the prior experiments and its corresponding multiple-choice question sets (used for the practice and criterial tests in counterbalanced fashion).

The procedure for the case of pretesting and conventional posttesting was similar to the multiple-choice/no-feedback conditions of Experiment 3 (see Figure 3), except that participants read the text passage twice in succession and that passage reading time was fixed at 8 minutes (which, given the doubled exposure to the text, was an added measure to increase experimental control). For the Read-Test-Read condition, the passage was first read once for 8 minutes, followed by a posttest on that passage, and then the passage was read a second time for 8 minutes (it should be noted that although the Read-Test-Read condition bears a resemblance to the interpolated testing methods used in some prior studies, in the present case there was only a single practice test rather than multiple tests). Overall, participants in all three conditions completed two readings of the passage and a single 8-question multiple-choice practice test, with only the order of those activities differing. After a 5-min distractor task, the criterial test was administered.

Results and Discussion

Learning Phase

Practice test performance was greater in the case of conventional posttesting (M = 55%, SD = 24%) and in the Read-Test-Read condition (M = 49%, SD = 23%) than pretesting (M = 30%, SD = 17%), which generally matches the patterns observed in the prior experiments.

Criterial Test

Control questions. Performance on control questions was similar in the pretesting (M = 48%, SD = 32%), conventional posttesting (M = 49% SD = 26%), and Read-Test-Read conditions (M = 48%, SD = 28%), which is similar to the patterns observed in the prior experiments.

We conducted an ANOVA on criterial test scores (see Table 3) with factors of Test Type (Pretest vs. Posttest vs. Read-Test-Read) and Question Type (Tested vs. Untested). There was a significant main effect of Test Type, indicating that performance on the criterial test differed between conditions, $F(2, 395) = 17.55$, $p < .0001$, $\cdot_p^2 = 0.08$. The main effect of Question Type was not significant, indicating that performance was similar for Tested and Untested questions, $F(1, 395) = 1.44$, $p = .23$. There was also a significant interaction between Test Type and Question Type, $F(2, 395) = 13.10$, $p < .0001$, $\cdot_p^2 = 0.06$. For the case of Tested questions, criterial test scores were substantially higher following pretesting than conventional posttesting, $t(267) = 5.15$, $p < .0001$, $d = 0.62$, and higher in the Read-Test-Read condition than in the conventional posttesting condition, $t(253) = 7.07$, $p < .0001$, $d = 0.88$, but not significantly different between the pretesting and Read-Test-Read conditions, $t(269) = -1.77$, $p = .078$. For the case of Untested questions, criterial test scores were also higher following pretesting than conventional posttesting, $t(267) = 3.15$, $p = .0018$, $d = 0.38$, and higher in the Read-Test-Read condition than in the conventional posttesting condition, $t(253) = 3.14$, $p = .0019$, $d = 0.39$, but

not significantly different between the pretesting and Read-Test-Read conditions, $t(268) = 0.0076$, $p = .99$. Further, as indicated by the significant Test Type by Question Type interaction, the advantage of the pretesting and Read-Test-Read conditions over the conventional posttesting condition was smaller for Untested than Tested questions. Overall, the advantage of pretesting over conventional posttesting (percent increase) was $M = 27\%$ and $M = 11\%$ for Tested and Untested questions, respectively, and the corresponding advantage of the Read-Test-Read condition over conventional posttesting was $M = 38\%$ and $M = 13\%$ for Tested and Untested questions, respectively. These patterns are consistent with the likelihood that test-potentiated learning was responsible for the pretesting advantage in the present experiments. Specifically, performance suffered when participants were denied an opportunity to read the text passage after taking a practice test, as occurred with conventional implementations of posttesting, and improved when such an opportunity was given, as in the case of pretesting and in the Read-Test-Read condition of Experiment 5.

Supplementary Meta-Analyses of Experiments 1-4

Relative Efficacy of Pretesting Versus Posttesting

For further insights into the pretesting advantages observed across experiments, we conducted two internal meta-analyses of the results from Experiments 1-4. Given design differences, and most notably the second reading opportunity, Experiment 5 was not included. These meta-analyses, which were performed separately for Tested and Untested questions, involved pretesting minus posttesting effect sizes in terms of Cohen's *d* (i.e., the effect size derived from a *t*-test comparing pretest and posttest performance), wherein a positive *d*-value represented an advantage of pretesting over posttesting and a negative *d*-value represented the reverse case. The sampling variance for each effect size, *sv*, was calculated using equations for within-subjects and between-subjects designs as specified in Morris and DeShon (2002, p. 117).

The Cohen's *d* and *sv* values were entered into the *metafor* package in R (Viechtbauer, 2010) and random-effects meta-analyses were performed. Although the meta-analyses reported here all involved data from the same participants and the same study, the existence of dependencies between those meta-analyses do not constitute a violation of the assumption of independence within meta-analysis (for discussion see Goh et al., 2016; Marín-Martínez & Sánchez-Meca, 1999); separate meta-analyses were conducted for each outcome measure and each meta-analysis was interpreted on its own (i.e., the conclusions do not rely on comparisons across internal meta-analyses).

As illustrated in the forest plot in Figure 3, there were 22 comparisons of pretesting versus posttesting involving Tested or Untested questions across Experiments 1-4. Pretesting increased performance by $d = 0.30$, $p < .001$, 95% CI [0.17, 0.44] for Tested questions and $d = 0.14$, $p = .0002$, 95% CI [0.05, 0.23] for Untested questions. Those analyses reinforce the conclusion that pretesting yielded significantly better retention and, in most cases, also transfer of learning.

Pretesting and Testing Effects

Using the same meta-analytic procedures, we also computed effect size estimates for the pretesting and testing effects, as well as corresponding transfer effects, relative to performance on Control questions (which in these analyses served as a non-testing reference condition). A forest plot of the resulting pretesting and testing effects is displayed in Figure 4 and a corresponding forest plot of transfer effects is displayed in Figure 5.

For Tested versus Control questions, the pretesting effect was $d = 1.02$, $p < .001$, 95% CI [0.70, 1.33], and the testing effect was $d = 0.76$, $p < .001$, 95% CI [0.46, 1.06]. Both effects are somewhat larger than the testing effect of $g = 0.50$ reported by Rowland (2014; cf. Adesope, 2017), although the testing effect in that meta-analysis were calculated relative to a more

stringent reexposure control (e.g., restudying) and hence might be expected to be smaller. For Untested versus Control questions, transfer of learning was estimated at $d = 0.20$, $p < .001$, 95% CI [0.08, 0.32] following pretesting and $d = 0.09$, $p = .018$, 95% CI [0.01, 0.16] following posttesting. Relatedly, Pan and Rickard (2018) reported a similar effect size estimate for posttesting and transfer to untested text passage materials, albeit relative to a more stringent reexposure control ($d = 0.16$, 95% CI [-0.10, 0.43]).

General Discussion

Across five experiments, we investigated the relative efficacy of pretesting and posttesting at enhancing learning. A pretesting advantage was repeatedly observed: Pretesting yielded higher criterial test scores for Tested questions in all five experiments, and higher criterial test scores for Untested questions in all but Experiment 4. These patterns were robust to different variants of practice testing and other aspects of experimental design, including the use of multiple-choice or cued recall tests, the presence or absence of correct answer feedback, 5-minute or 48-hour retention intervals, online and undergraduate student participants, and whether participation occurred remotely or in a supervised lab setting. Supplementary meta-analyses further revealed that both pretesting and posttesting enhanced learning relative to a no-testing control condition: Both test types significantly improved memory, but pretesting did so to a greater extent. Overall, these results reveal that pretesting can be highly competitive with posttesting and can yield similar, and in some cases greater, pedagogical benefits.

To our knowledge, the present study is the first to show such a consistent benefit of pretesting over posttesting. Although that finding might seem at odds with the handful of prior studies that included similar comparisons and involved the learning of educationally-relevant materials, some of which found evidence favoring posttesting (e.g., Latimier et al., 2019; McDaniel et al., 2011; Rothkopf & Bibiscos, 1967), multiple design features differentiate this

study from prior research. These differences include the use of a single practice test per passage rather than interpolated tests, the lack of feedback in most conditions, no reading opportunities after posttesting in most experiments, and the fact that participants did not typically score above chance on the pretests (which indicates minimal influence of prior knowledge). The testing procedures used in the present study also resembled common practices in the pretesting and posttesting literatures, thus facilitating an arguably simpler and more direct comparison of the two test types than in prior research. Under those circumstances, a significant pretesting advantage over posttesting (*d*s = 0.30 and 0.14 for Tested and Untested questions, respectively, in Experiments 1-4, and *d*s = 0.62 and 0.38 for Tested and Untested questions, respectively, in Experiment 5, not including the Read-Test-Read condition) was observed.

Why Is Pretesting Competitive with Posttesting?

Although the present study was not intended as an extensive investigation of the cognitive mechanisms that pretesting and posttesting engage (for in-depth discussions of the former see Anderson & Biddle, 1975; Hamaker, 1986; Kornell & Vaughn, 2016; Metcalfe, 2017; for discussions of the latter see Karpicke et al., 2014; Rowland, 2014, van den Broek et al., 2016), consideration of potential differences in the mechanisms that the test types engage—and especially the results of Experiment 5—enables us to propose the following tentative theoretical account. First consider posttesting. In Experiments 1-4, posttesting always involved reading a text passage and then taking a single practice test on it. Given that sequence of events, posttesting could not have impacted reading behavior. Rather, posttesting likely enhanced memory for passage content, and did so to the extent that such content was well-encoded during reading *and* successfully retrieved during the posttest (and when it was provided, feedback enhanced learning in cases of unsuccessful retrieval). Posttesting thus helped consolidate prior learning and was efficacious at doing so relative to a non-testing control condition. It was only

when a second reading opportunity occurred after a posttest, as was provided in the Read-Test-Read condition of Experiment 5, did an implementation of posttesting (that is, sandwiched between two reading opportunities) become as efficacious as pretesting. That finding suggests that a lack of opportunities for test-potentiated learning, as occurs in many conventional implementations of posttesting, can be disadvantageous for learning.

In contrast, the advantage of pretesting in the present experiments appeared to stem from test-potentiated learning. That conclusion is implied by the results of Experiment 5, wherein the addition of a second reading opportunity after testing eliminated the posttesting deficit relative to the pretesting condition. In the pretesting literature, pretesting has been hypothesized to impact subsequent reading or studying behaviors in a variety of possible ways (e.g., fostering greater overall levels of attention or inducing a search for correct answers; for discussions see Rickards, 1977; Geller et al., 2017; St. Hilaire & Carpenter, 2020). Although there were no significant reading time differences in nearly all cases, effects of pretesting on other reading behaviors, such as improved attention, are plausible in our view, and that improved attention may have resulted in test-potentiated learning. In support of the attentional explanation, Pan et al. (2020b) recently used mind-wandering probes to demonstrate that pretesting improves attentional focus to video lectures, relative to conditions wherein no pretests occur. Beyond simply improved attention, other cognitive processes, such as focusing on content that was emphasized during pretesting, might also lead to test-potentiated learning.

The test-potentiated learning that results from pretesting appears to be analogous to the effects of organizational signals (such as titles, headings, typographical cues, preview sentences, and other signaling devices) on text processing. In the literature on organizational signals (for a review, see Lorch, 1989), virtually all such types of signals improve memory for information that is cued in a text (e.g., memory for passage content cued with headings is improved relative to

memory for the same passage lacking such headings). The effects of organizational signals on memory may stem from changes in attention, reading behavior, and other cognitive processes. In the absence of such signals, however, learners may engage in a "default" or unfocused approach when reading text (Lorch, 1995). We suspect that such an unfocused approach also occurred in the posttesting condition in the present experiments, but not in the pretesting condition, which experienced test-potentiated learning as a result. Relatedly, Richland et al. (2009) demonstrated that two types of organizational signals, namely bolding and italics, were less effective at enhancing learning from text passages than pretesting, and also found that the combination of bolding/italics and pretesting was more effective than bolding/italics alone (see also Golding & Fowler, 1992). Thus, the effects of pretesting and organizational signals appear to be complementary, and pretesting can yield even more potent effects on learning than some types of organizational signals.

Overall, it is likely that the pretesting advantage in the present experiments resulted not directly from the act of testing itself, but instead from enhanced encoding of the text passage that followed. That conclusion is substantiated by the results of Experiment 5 and the relative underperformance of posttesting, in all five experiments, when an opportunity to read the text passage after practice testing was not provided. It should be emphasized that although test-potentiated learning is a typical consequence of pretesting, such learning is not exclusive to that test type; the provision of study opportunities after practice testing can also yield similar results following posttesting. Test-potentiated learning might also have been relevant in prior studies wherein interpolated posttesting was as effective or more effective than interpolated pretesting (e.g., Rothkopf & Bibiscos, 1967), given that such studies included reading opportunities after posttest questions had been attempted (although commonly for new content, which may have yielded test-potentiated learning of that content; for discussion see Chan et al., 2018).

Moderators of the Pedagogical Benefits of Pretesting and Posttesting

Across all experiments, we observed larger benefits of pretesting and posttesting for Tested as opposed to Untested questions. That result is consistent with the finding that both test types tend to yield greater retention than transfer of learning (Pan & Rickard, 2018; see also Carpenter et al., 2017; Hausman & Rhodes, 2018; James & Storm, 2019; and Toftness et al., 2017). In the present experiments, positive transfer relative to a no-testing control was observed, but the magnitude of that transfer was relatively small. Relatedly, some researchers have theorized that pretesting is more likely than posttesting to direct attention away from untested to tested information, thus reducing the likelihood of successful transfer (e.g., Frase, 1967; Sagaria & Di Vesta, 1977; see also Hannafin & Hughes, 1986; Wager & Wager, 1985), but our results are not consistent with that suggestion. Overall, the present experiments reinforce the conclusion that both test types are more likely to yield benefits on measures of retention than transfer.

The lack of a significant effect of test format in Experiments 2 and 3 is consistent with the finding that test format is not a strong moderator of the testing effect (Pan & Rickard, 2018; Smith & Karpicke, 2014; cf. Rowland, 2014), and further reinforces the conclusion that pretesting is equally or more effective than posttesting across a variety of circumstances. In Experiment 4, however, a benefit of multiple-choice over cued recall testing was observed for the case of pretesting and posttesting without feedback, which suggests that a multiple-choice advantage may be more apparent after an extended retention interval. Additionally, transfer to Untested questions was greater following multiple-choice versus cued recall testing in Experiment 4 but not in Experiments 2 and 3. Little et al. (2012) and Little and Bjork (2016) have theorized that multiple-choice tests with competitive answer alternatives can facilitate more successful transfer than cued recall tests and especially when instructions to carefully consider each answer alternative or engage in extended deliberation are used (Little, 2011; Experiment 5).

Our participants were told to read each question carefully and to select the best answer, but not asked to engage in extended deliberation; the lack of such instructions may have attenuated some of the benefits of multiple-choice tests.

Immediate correct answer feedback enhanced performance for Tested questions in Experiments 3 and 4, which aligns with results from the posttesting literature (e.g., Rowland, 2014) and suggests that adding feedback to pretests can enhance learning (cf. Sana et al., in press). Further, the fact that feedback did not impair the pretesting benefit, at least for Tested questions, challenges theoretical accounts which attribute the pretesting effect to a search for correct answers (but does not rule out the possibility that pretesting enhances attention or other learning behaviors). With respect to Untested questions, however, feedback had no effect on the efficacy of pretesting and a negative effect on posttesting in Experiment 4. As previously noted, such feedback was directly relevant for Tested questions and not for Untested questions (wherein feedback involved the same information category but otherwise entailed competing information). Hence, it is not altogether surprising that feedback was not beneficial for performance on Untested questions.

The source of potentially deleterious effects of feedback on recall of previously untested information in Experiment 4, however, remains to be determined. One possibility is that correct answer feedback, which entailed the answers to Tested questions only, focused participants' attention on content targeted by Tested questions at the expense of content targeted by Untested questions (i.e., the other multiple-choice answer options, when provided); absent such feedback, no such focusing occurred. A second consideration involves the lesser background knowledge of the undergraduate participants in Experiment 4 relative to the MTurk participants of the prior experiments (as indicated by practice test performance; for related evidence see Casler et al., 2013; Hauser & Schwarz, 2015). Having lower baseline knowledge may reduce the likelihood

of successful transfer. It is important however to emphasize that our data do not indicate that correct answer feedback necessarily always has negative effects on the recall of Untested information (Experiments 3 vs. 4), and positive transfer in the absence of feedback was observed among the undergraduate participants in Experiment 5.

Limitations

A major limitation of the present research is that all experiments involved the same set of materials, namely encyclopedia-style expository texts, and in which the target information were multiple distinct categories of information with at least four exemplars per category. Moreover, for the case of multiple-choice questions, the answer options for a given question all drew from the same category and were plausible or competitive alternatives with one another (cf. Little, 2011). As such, although the text passages used in the present study resemble those found in textbook chapters and other learning contexts, it remains to be determined whether the same results would be obtained if any aspect of the materials were changed, such as to the science lessons used by McDaniel et al. (2011) and Latimier et al. (2019; for the case of fill-in-the-blank vocabulary words, see de Lima & Jaeger, in press), or if the materials being learned were not category exemplars for which the correct answers were explicitly stated in the text (e.g., generating inferences from the text).

Another limitation is that our participants were not incentivized to learn the target materials to an especially high level of mastery. If successful completion of the study was contingent on performance, then it is possible that more learning would have occurred; alternatively, such pressure may have had deleterious consequences. Further, we compared both test types against one another and not against an exposure-matched control (e.g., restudying); such a condition would have provided a measure of the relative efficacy of engaging in a non-testing activity (although, as previously noted, prior studies have included such controls). Other

limitations involve the mode of presentation, namely via computer, question-by-question, and without the ability to review one's answers; an in-class implementation of pretesting or posttesting might differ substantially from those procedures. More broadly, although we submit that similar results are likely to obtained under comparable circumstances—that is, with other expository texts, similar types of practice and criterial test questions, and without strong incentives to learn (as may occur with low-stakes practice tests)—the generalization of our findings to all other uses of test-enhanced learning should be made cautiously and pending further research.

Future Directions

Future research stands to yield further insights into the relative benefits of pretesting and posttesting. One possibility is that presentation modality, which was not manipulated in the present study, moderates the relative benefits of the two test types (and particularly whether positive transfer of learning occurs). Carpenter and Toftness (2017) have argued that specific benefits of pretesting (i.e., lack of transfer) are more likely if learners study information in the form of expository texts, with learners selectively attending to parts of the text that are deemed important and engaging in shallow processing of untested information. Using video materials (which are not necessarily self-paced and may reduce the likelihood of skipping through information) could address that issue; indeed, pretesting with video materials can yield greater amounts of transfer (Carpenter & Toftness, 2017; cf. James & Storm, 2019).

Future studies of pretesting and posttesting might also use application, inference, or problem-solving questions as measures of transfer, as well as higher- or lower-order practice questions. Rickards (1977), Jensen et al. (2014), and others have argued that higher-order questions are more effective at yielding transferrable learning, although pretest questions that require generating inferences were not especially effective in one recent study (Hausman &

Rhodes, 2018). Moreover, whether the advantage of pretesting over posttesting is typically limited to directly tested content, as was observed at a 48-hour interval in Experiment 4, and the potential role of feedback for transfer, needs to be further investigated. Such research might use materials that can be delineated into question-relevant and -irrelevant information.

Test-potentiated learning also needs to be explored further in order to better clarify the circumstances under which such learning manifests, the cognitive processes involved, and its effects on the efficacy of practice testing. As one example, recent research on retrieval-enhanced suggestibility indicates that posttesting increases susceptibility to encoding subsequently presented false information via test-potentiated learning (LaPaglia & Chan, 2019; see also Manley & Chan, 2019). Finally, the effects of combining pretesting and posttesting (cf. Carpenter et al., 2018), and the relative benefits of non-interpolated, interpolated, and repeated practice tests warrant further research.

Broader Implications

The present results challenge the notion that retrieval practice is always more pedagogically potent than errorful generation, which currently seems to be the prevailing view, both empirically and anecdotally. That perception possibly stems from the fact that, with some exceptions, research on test-enhanced learning has focused largely on the benefits of posttesting, and substantially less attention has been devoted to the potential benefits of pretesting. The current experiments help address this gap. Our findings suggest that pretesting can be as potent as posttesting, if not more so, regardless of test format and the presence or absence of feedback. Accordingly, an individual wishing to use test-enhanced learning for a text passage, book chapter, or other similar types of materials could justifiably consider both test types. Although a recommendation to incorporate pretesting instead of posttesting in lectures and study sessions is premature at this point, and the efficacy of both test types for learning other kinds of materials

need to be directly compared, it is now increasingly evident that retrieval practice is not the only viable type of practice testing. Indeed, students and instructors would be well-advised to consider including both pretesting and posttesting in their learning repertoire.

References

Abbott, E. E. (1909). On the analysis of the factors of recall in the learning process. *Psychological Monographs.* 11, 159–177. doi: 10.1037/h0093018

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. doi: 10.3102/0034654316689306

Anderson, R. C., & Biddle, W. B. (1975). On asking people questions about what they are reading. In *Psychology of Learning and Motivation* (Vol. 9, pp. 89-132). Academic Press. doi: 10.1016/S0079-7421(08)60269-8

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

Barnett, S. M., & Ceci, S. J. (2002). When and where do we apply what we learn?: A taxonomy for far transfer. *Psychological Bulletin*, *128*(4), 612-637. doi: 10.1037//0033-2909.128.4.61

Bjork, R. A. (1975). Retrieval as a memory modifier: An interpretation of negative recency and related phenomena. In R. L. Solso (Ed.), *Information processing and cognition: The Loyola Symposium* (pp. 123–144). Hillsdale, NJ: Erlbaum.

Bjork, E. L., & Bjork, R. A. (2011). Making things hard on yourself, but in a good way: Creating desirable difficulties to enhance learning. *Psychology and the real world: Essays illustrating fundamental contributions to society*, *2*(59-68).

Bjork, R. A., Dunlosky, J., & Kornell, N. (2013). Self-regulated learning: Beliefs, techniques, and illusions. *Annual Review of Psychology, 64*, 417-444. doi: 10.1016/S0079-7421(08)60269-8

Brown, P. C., Roediger III, H. L., & McDaniel, M. A. (2014). *Make it stick*. Harvard University Press.

Butterfield, B., & Metcalfe, J. (2001). Errors committed with high confidence are hypercorrected. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 27*(6), 1491-1494. doi: 10.1016/S0079-7421(08)60269-8

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. doi: 10.1037/a0017021

Carpenter, S. K., Pashler, H., & Cepeda, N. J. (2009). Using tests to enhance 8th grade students' retention of US history facts. *Applied Cognitive Psychology, 23*(6), 760-771. doi: 10.1002/acp.1507

Carpenter, S. K., Pashler, H., Wixted, J. T., & Vul, E. (2008). The effects of tests on learning and forgetting. *Memory & Cognition, 36*(2), 438-448. doi: 10.3758/MC.36.2.438

Carpenter, S. K., Rahman, S., & Perkins, K. (2018). The effects of prequestions on classroom learning. *Journal of Experimental Psychology: Applied, 24*(1), 34-42. doi: 10.1037/xap0000145

Carpenter, S. K., & Toftness, A. R. (2017). The effect of prequestions on learning from video presentations. *Journal of Applied Research in Memory and Cognition, 6*(1), 104-109. doi: 10.1016/j.jarmac.2016.07.014

Casler, K., Bickel, L., & Hackett, E. (2013). Separate but equal? A comparison of participants and data gathered via Amazon’s MTurk, social media, and face-to-face behavioral testing. *Computers in Human Behavior, 29*(6), 2156–2160. doi: 10.1016/j.chb.2013.05.009

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

Cheung, M. W.-L. (2019). A guide to conducting a meta-analysis with non-independent effect sizes. *Neuropsychology Review*, *29*(4), 387–396. doi: 10.1007/s11065-019-09415-6

Davis, S. D., Chan, J. C., & Wilford, M. M. (2017). The dark side of interpolated testing: Frequent switching between retrieval and encoding impairs new learning. *Journal of Applied Research in Memory and Cognition*, *6*(4), 434-441. doi: 10.1016/j.jarmac.2017.07.002

de Lima, N. K., & Jaeger, A. (2020). The Effects of Prequestions versus Postquestions on Memory Retention in Children. *Journal of Applied Research in Memory and Cognition* (online first publication). doi: 10.1016/j.jarmac.2020.08.005

Dunlosky, J., Rawson, K. A., Marsh, E. J., Nathan, M. J., & Willingham, D. T. (2013). Improving students' learning with effective learning techniques: Promising directions from cognitive and educational psychology. *Psychological Science in the Public Interest, 14*(1), 4-58. doi: 10.1177/1529100612453266

Faul, F., Erdfelder, E., Lang, A. G., & Buchner, A. (2007). G* Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*(2), 175-191. doi: 10.3758/BF03193146

Frase, L. T. (1967). Learning from prose material: Length of passage, knowledge of results, and position of questions. *Journal of Educational Psychology*, *58*(5), 266-272. doi: 10.1037/h0025028

Frase, L. T. (1968). Effect of question location, pacing, and mode upon retention of prose material. *Journal of Educational Psychology*, *59*(4), 244-249. doi: 10.1037/h0025947

Geller, J., Toftness, A. R., Armstrong, P. I., Carpenter, S. K., Manz, C. L., Coffman, C. R., & Lamm, M. H. (2018). Study strategies and beliefs about learning as a function of academic achievement and achievement goals. *Memory, 26*(5), 683-690. doi: 10.1037/h0025947

Gelman, A. (2018, March 15). You need 16 times the sample size to estimate an interaction than to estimate a main effect. *Statistical Modeling, Casual Inference, and Social Science.* https://statmodeling.stat.columbia.edu/2018/03/15/need-16-times-sample-size-estimate-interaction-estimate-main-effect/

Goh, J. X., Hall, J. A., & Rosenthal, R. (2016). Mini meta-analysis of your own studies: Some arguments on why and a primer on how. *Social and Personality Psychology Compass*, *10*(10), 535–549. doi: 10.1111/spc3.12267

Golding, J. M., & Fowler, S. B. (1992). The limited facilitative effect of typographical signals. *Contemporary Educational Psychology, 17*(2), 99-113. doi: 10.1016/0361-476X(92)90052-Z

Hannafin, M. J., & Hughes, C. W. (1986). A framework for incorporating orienting activities in computer-based interactive video. *Instructional Science, 15*(1), 239-255. doi: 10.1007/BF00139613

Hamaker, C. (1986). The effects of adjunct questions on prose learning. *Review of Educational Research*, *56*(2), 212-242. doi: 10.3102/00346543056002212

Hartwig, M. K., & Dunlosky, J. (2012). Study strategies of college students: Are self-testing and scheduling related to achievement?. *Psychonomic Bulletin & Review*, *19*(1), 126-134. doi: 10.3758/s13423-011-0181-y

Hauser, D. J., & Schwarz, N. (2016). Attentive Turkers: MTurk participants perform better on online attention checks than do subject pool participants. *Behavior Research Methods, 48*(1), 400–407. doi: 10.3758/s13428-015-0578-z

Hausman, H., & Rhodes, M. G. (2018). When pretesting fails to enhance learning concepts from reading texts. *Journal of Experimental Psychology: Applied, 24*(3), 331-346. doi: 10.1037/xap0000160

Hays, M. J., Kornell, N., & Bjork, R. A. (2013). When and why a failed test potentiates the effectiveness of subsequent study. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 39*(1), 290-296. doi: 10.1037/a0028468

Hinze, S. R., & Rapp, D. N. (2014). Retrieval (sometimes) enhances learning: Performance pressure reduces the benefits of retrieval practice. *Applied Cognitive Psychology, 28*(4), 597-606. doi: 10.1002/acp.3032

Jensen, J. L., McDaniel, M. A., Woodard, S. M., & Kummer, T. A. (2014). Teaching to the test… or testing to teach: Exams requiring higher order thinking skills encourage greater conceptual understanding. *Educational Psychology Review, 26*(2), 307-329. doi: 10.1007/s10648-013-9248-9

Kang, S. H., McDermott, K. B., & Roediger III, 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. doi: 10.1080/09541440601056620

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

Kornell, N. (2014). Attempting to answer a meaningful question enhances subsequent learning even when feedback is delayed. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 40*(1), 106-114. doi: 10.1037/a0033699

Kornell, N., & Bjork, R. A. (2007). The promise and perils of self-regulated study. *Psychonomic Bulletin & Review, 14*(2), 219-224. doi: 10.3758/BF03194055

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. doi: 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. doi: 10.1037/a0015729

Kornell, N., & Vaughn, K. E. (2016). How retrieval attempts affect learning: A review and synthesis. In *Psychology of learning and motivation* (Vol. 65, pp. 183-215). Academic Press. doi: 10.1016/bs.plm.2016.03.003

LaPaglia, J. A., & Chan, J. C. (2019). Telling a good story: The effects of memory retrieval and context processing on eyewitness suggestibility. *PloS one, 14*(2), e0212592. doi: 10.1371/journal.pone.0212592

Latimier, A., Riegert, A., Peyre, H., Ly, S. T., Casati, R., & Ramus, F. (2019). Does pre-testing promote better retention than post-testing?. *NPJ science of learning, 4*(1), 1-7. doi: 10.1038/s41539-019-0053-1

Limesurvey GmbH. LimeSurvey: An Open Source survey tool. LimeSurvey GmbH, Hamburg, Germany. URL: http://www.limesurvey.org

Little, J. L. (2011). Optimizing multiple-choice tests as learning events (Order No. AAI3493389). doi: 10.3758/s13421-016-0621-z

Little, J. L., & Bjork, E. L. (2016). Multiple-choice pretesting potentiates learning of related information. *Memory & Cognition, 44*(7), 1085-1101. doi: 10.3758/s13421-016-0621-z

Little, J. L., Bjork, E. L., Bjork, R. A., & Angello, G. (2012). Multiple-choice tests exonerated, at least of some charges: Fostering test-induced learning and avoiding test-induced forgetting. *Psychological Science, 23*(11), 1337-1344. doi: 10.1177/0956797612443370

Lorch, R. F. (1989). Text-signaling devices and their effects on reading and memory processes. *Educational Psychology Review, 1*(3), 209-234. doi: 10.1007/BF01320135

Lorch Jr, R. F., & Lorch, E. P. (1995). Effects of organizational signals on text-processing strategies. *Journal of Educational Psychology, 87*(4), 537-544. doi: 10.1037/0022-0663.87.4.537

Manley, K. D., & Chan, J. C. K. (2019). Does Retrieval Enhance Suggestibility Because It Increases Perceived Credibility of the Postevent Information? *Journal of Applied Research in Memory and Cognition, 8*(3), 355–366. doi: 10.1016/j.jarmac.2019.06.001

Marín-Martínez, F., & Sánchez-Meca, J. (1999). Averaging dependent effect sizes in meta-analysis: A cautionary note about procedures. *The Spanish Journal of Psychology, 2*, 32–38. doi: 10.1017/S1138741600005436

McCrudden, M. T., & Schraw, G. (2007). Relevance and goal-focusing in text processing. *Educational Psychology Review, 19*(2), 113-139. doi: 10.1007/s10648-006-9010-7

McCrudden, M. T., Schraw, G., & Kambe, G. (2005). The Effect of Relevance Instructions on Reading Time and Learning. *Journal of Educational Psychology, 97*(1), 88-102. doi: 10.1037/0022-0663.97.1.88

McDaniel, M. A., Agarwal, P. K., Huelser, B. J., McDermott, K. B., & Roediger III, H. L. (2011). Test-enhanced learning in a middle school science classroom: The effects of quiz frequency and placement. *Journal of Educational Psychology*, *103*(2), 399-414. doi: 10.1037/a0021782

Metcalfe, J., & Huelser, B. J. (2020). Learning from errors is attributable to episodic recollection rather than semantic mediation. *Neuropsychologia*, *138*, 107296. doi: 10.1016/j.neuropsychologia.2019.107296

Meyer, A. N., & Logan, J. M. (2013). Taking the testing effect beyond the college freshman: Benefits for lifelong learning. *Psychology and Aging*, *28*(1), 142-147. doi: 10.1037/a0030890

Microsoft Corporation (2011). Microsoft Fuzzy Lookup Add-In for Excel. https://atidan.files.wordpress.com/2013/08/fuzzy-lookup-add-in-for-excel.pdf

Morris, S. B., & DeShon, R. P. (2002). Combining effect size estimates in meta-analysis with repeated measures and independent-groups designs. *Psychological Methods*, *7*(1), 105-125. doi: 10.1037/1082-989X.7.1.105

Pan, S.C., & Agarwal, P.K. (2018). *Retrieval practice and transfer of learning: Fostering students' application of knowledge.* San Diego, CA: UC San Diego. Retrieved from http://pdf.retrievalpractice.org/TransferGuide.pdf

Pan, S. C., Cooke, J., Little, J. L., McDaniel, M. A., Foster, E. R., Connor, L. T., & Rickard, T. C. (2019). Online and clicker quizzing on jargon terms enhances definition-focused but not conceptually focused biology exam performance. *CBE—Life Sciences Education*, *18*(4), ar54.

Pan, S. C., Hutter, S. A., D'Andrea, D., Unwalla, D., & Rickard, T. C. (2019). In search of transfer following cued recall practice: The case of process•based biology concepts. *Applied Cognitive Psychology*, *33*(4), 629-645.

Pan, S. C., Lovelett, J., Stoeckenius, D., & Rickard, T. C. (2019). Conditions of highly specific learning through cued recall. *Psychonomic Bulletin & Review*, *26*(2), 634-640. doi: 10.3758/s13423-019-01593-x

Pan, S. C., & Rickard, T. C. (2017). Does retrieval practice enhance learning and transfer relative to restudy for term-definition facts?. *Journal of Experimental Psychology: Applied*, *23*(3), 278-292. doi: 10.1037/xap0000124

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

Pan, S. C., Sana, F., Samani, J., Cooke, J., and Kim, J. A. (2020a). Learning from errors: Students' and instructors' practices, attitudes, and beliefs. *Memory* (online first publication). doi: 10.1080/09658211.2020.1815790

Pan, S. C., Sana, F., Schmitt, A., and Bjork, E. L. (2020b). Pretesting reduces mind wandering and enhances learning during online lectures. *Journal of Applied Research in Memory and Cognition* (online first publication). doi: 10.1016/j.jarmac.2020.07.004

Pashler, H., Bain, P. M., Bottge, B. A., Graesser, A., Koedinger, K., McDaniel, M., & Metcalfe, J. (2007). Organizing Instruction and Study to Improve Student Learning. IES Practice Guide. NCER 2007-2004. *National Center for Education Research*. doi: 10.1037/e607972011-001

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

Pyc, M. A., & Rawson, K. A. (2010). Why testing improves memory: Mediator effectiveness hypothesis. *Science, 330*(6002), 335-335. doi: 10.1126/science.1191465

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. doi: 10.1037/a0023956

Richland, L. E., Kornell, N., & Kao, L. S. (2009). The pretesting effect: Do unsuccessful retrieval attempts enhance learning?. *Journal of Experimental Psychology: Applied*, *15*(3), 243-257. doi: 10.1037/a0016496

Rickard, T. C., & Pan, S. C. (2017). Time for considering the possibility that sleep plays no unique role in motor memory consolidation: Reply to Adi-Japha and Karni (2016). *Psychological Bulletin, 143*(4), 454–458. https://doi.org/10.1037/bul0000094

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

Rickards, J. P. (1976). Interaction of position and conceptual level of adjunct questions on immediate and delayed retention of text. *Journal of Educational Psychology, 68*(2), 210-217. doi: 10.1037/0022-0663.68.2.210

Rickards, J. P. (1977). On inserting questions before or after segments of text. *Contemporary Educational Psychology*, *2*(2), 200-206. doi: 10.1016/0361-476X(77)90021-2

Roediger III, H. L., & Butler, A. C. (2011). The critical role of retrieval practice in long-term retention. *Trends in Cognitive Sciences, 15*(1), 20-27. doi: 10.1016/j.tics.2010.09.003

Roediger III, H. L., & Karpicke, J. D. (2006). The power of testing memory: Basic research and implications for educational practice. *Perspectives on Psychological Science, 1*(3), 181-210. doi: 10.1111/j.1745-6916.2006.00012.x

Rothkopf, E. Z. (1966). Learning from written instructive materials: An exploration of the control of inspection behavior by test-like events. *American Educational Research Journal*, *3*(4), 241-249. doi: 10.3102/00028312003004241

Rothkopf, E. Z., & Bisbicos, E. E. (1967). Selective facilitative effects of interspersed questions on learning from written materials. *Journal of Educational Psychology*, *58*(1), 56-61. doi: 10.1037/h0024117

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. doi: 10.1037/a0037559

Sagaria, S. D., & Di Vesta, F. J. (1978). Learner expectations induced by adjunct questions and the retrieval of intentional and incidental information. *Journal of Educational Psychology, 70*(3), 280-288. doi: 10.1037/0022-0663.70.3.280

Sana, F., Forrin, N., Sharma, M., Dubljevic, T., Ho, P., Jalil, E., & Kim, J. A. (2020). Optimizing the efficacy of learning objectives through pretests. *CBE—Life Sciences Education* (in press).

Simonsohn, U. (2014, March 12). No-way interactions. *Data Colada.* http://datacolada.org/17 doi: 10.15200/winn.142559.90552

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

Soderstrom, N. C., & Bjork, R. A. (2015). Learning versus performance: An integrative review. *Perspectives on Psychological Science*, *10*(2), 176-199. doi: 10.1177/1745691615569000

St. Hilaire, K. J., & Carpenter, S. K. (2020). Prequestions enhance learning, but only when they are remembered. *Journal of Experimental Psychology: Applied.* Advance online publication. doi: 10.1037/xap0000296

St. Hilaire, K. J., Carpenter, S. K., & Jennings, J. M. (2019). Using prequestions to enhance learning from reading passages: the roles of question type and structure building ability. *Memory, 27*(9), 1204-1213. doi: 10.1080/09658211.2019.1641209

Szpunar, K. K., McDermott, K. B., & Roediger III, H. L. (2008). Testing during study insulates against the buildup of proactive interference. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *34*(6), 1392-1399. doi: 10.1037/a0013082

Swenson, I., & Kulhavy, R. W. (1974). Adjunct questions and the comprehension of prose by children. *Journal of Educational Psychology, 66*(2), 212-215. doi: doi.org/10.1037/h0036276

Toftness, A. R., Carpenter, S. K., Lauber, S., & Mickes, L. (2018). The limited effects of prequestions on learning from authentic lecture videos. *Journal of Applied Research in Memory and Cognition*, *7*(3), 370-378. doi: 10.1016/j.jarmac.2018.06.003

Van den Broek, G., Takashima, A., Wiklund-Hörnqvist, C., Wirebring, L. K., Segers, E., Verhoeven, L., & Nyberg, L. (2016). Neurocognitive mechanisms of the "testing effect": A review. *Trends in Neuroscience and Education*, *5*(2), 52-66. doi: 10.1016/j.tine.2016.05.001

Vaughn, K. E., & Rawson, K. A. (2012). When is guessing incorrectly better than studying for enhancing memory?. *Psychonomic Bulletin & Review*, *19*(5), 899-905. doi: 10.3758/s13423-012-0276-0

Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software, 36*(3), 1–48. doi: 10.18637/jss.v036.i03

Wager, W., & Wager, S. (1985). Presenting questions, processing responses, and providing feedback in CAI. *Journal of Instructional Development, 8*(4), 2-8. doi: 10.1007/BF02906047

Wissman, K. T., Rawson, K. A., & Pyc, M. A. (2011). The interim test effect: Testing prior material can facilitate the learning of new material. *Psychonomic Bulletin & Review, 18*(6), 1140-1147. doi: 10.3758/s13423-011-0140-7

Table 1

Experiments 1-4 Learning Phase Performance – Mean Percentage (SD) and Sample Size

Test Type	Test Format	Feedback	Experiment 1	Experiment 2	Experiment 3	Experiment 4
Pretest	Multiple choice	No	33 (18)	32 (19)	42 (19)	31 (13)
			174	139	61	34
		Yes	-	-	35 (15)	28 (15)
			-	-	67	29
	Cued recall	No		12 (16)	16 (23)	2 (4)
				134	58	29
		Yes	-	-	14 (18)	4 (7)
			-	-	64	32
Posttest	Multiple choice	No	61 (24)	62 (22)	71 (23)	57 (21)
			174	134	61	28
		Yes	-	-	65 (23)	57 (21)
			-	-	61	28
	Cued recall	No		46 (24)	47 (26)	26 (20)
				139	57	27
		Yes	-	-	46 (24)	24 (20)
			-	-	63	29

Table 2

Experiments 1-4 Criterial Test Performance – Mean Percentage (SD)

Test Type	Test Format	Feedback	Experiment 1			Experiment 2			Experiment 3			Experiment 4		
			Tested	Untested	Control	Tested	Untested	Control	Tested	Untested	Control	Tested	Untested	Control
Pretest	Multiple choice	No	69 (24)	61 (24)	56 (29)	70 (23)	59 (22)	56 (28)	73 (25)	57 (27)	50 (30)	75 (16)	49 (21)	38 (26)
		Yes	-	-	-	-	-	-	88 (16)	56 (25)	46 (28)	82 (16)	41 (21)	34 (30)
	Cued recall	No	-	-	-	69 (24)	59 (23)	57 (27)	77 (22)	64 (21)	47 (23)	63 (21)	37 (19)	34 (30)
		Yes	-	-	-	-	-	-	93 (11)	56 (25)	56 (28)	87 (13)	38 (20)	38 (23)
Posttest	Multiple choice	No	62 (24)	57 (25)	55 (30)	64 (21)	59 (21)	58 (29)	70 (22)	57 (30)	59 (27)	53 (19)	48 (21)	32 (26)
		Yes	-	-	-	-	-	-	88 (17)	55 (30)	54 (28)	72 (22)	42 (22)	37 (28)
	Cued recall	No	-	-	-	65 (22)	54 (24)	53 (30)	65 (22)	56 (25)	50 (28)	52 (18)	44 (19)	33 (25)
		Yes	-	-	-	-	-	-	91 (12)	52 (26)	52 (30)	82 (18)	31 (16)	33 (31)

Table 3

Experiment 5 Criterial Test Performance – Mean Percentage (SD)

Test Type and Condition	Tested	Untested	Control
Pretest	71 (27)	69 (23)	48 (32)
Posttest	56 (23)	62 (21)	49 (26)
Read-Test-Read	77 (25)	70 (22)	48 (28)

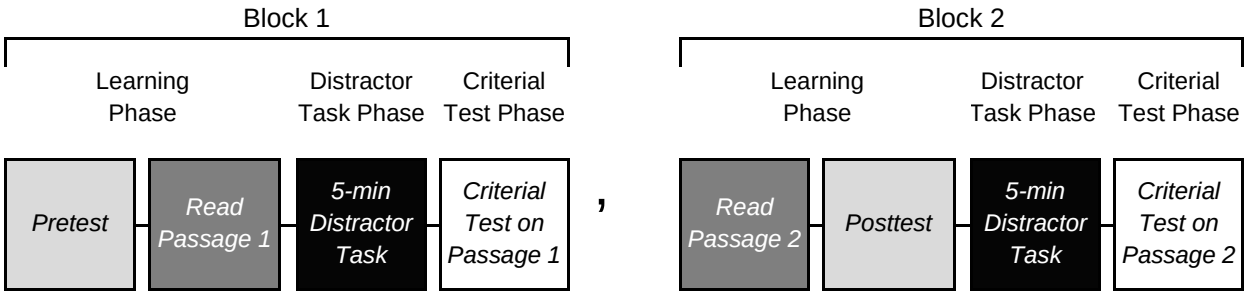

Figure 1

Overview of the experimental procedure for Experiments 1-2, and with some modifications, Experiments 3-4. In Experiments 1-2, participants completed two blocks, each involving a pretest or a posttest, the reading of a text passage, a 5-minute distractor task, and a criterial test. Experiments 3-4 involved the same procedure except that there was only one block and the criterial test was delayed by 48 hours in Experiment 4. For simplicity, one of two counterbalanced orders involving the placement of the pretested versus posttested passages is shown.

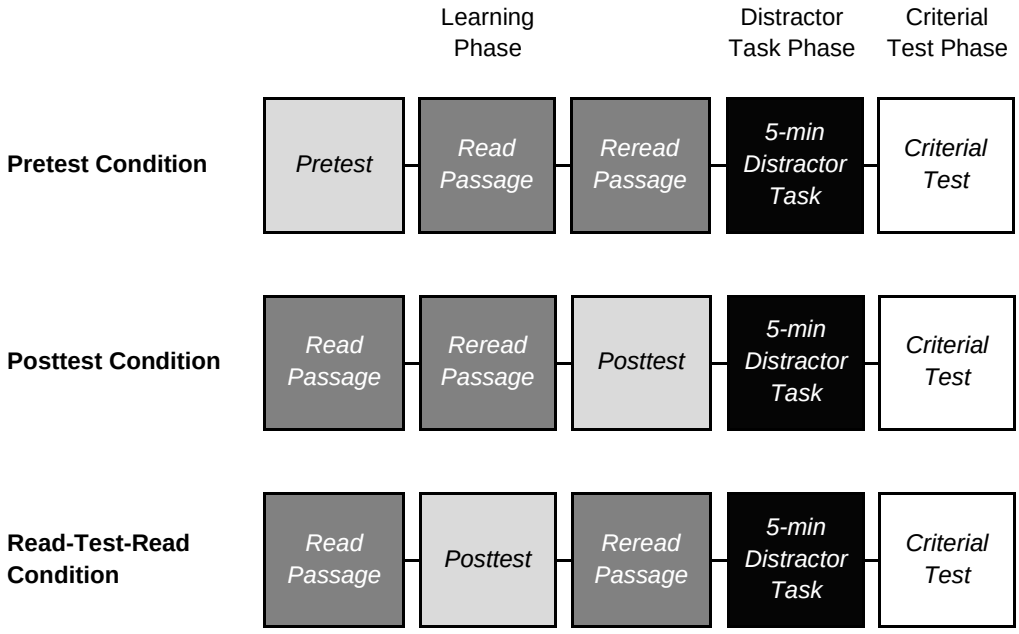

Figure 2

Overview of the procedure for Experiment 5. Participants completed a single training block in which they took a pretest or posttest and read a text passage twice. There were three between-subjects conditions. In the *Pretest* condition, the pretest occurred prior to reading the passage. In the (conventional) *Posttest* condition, the posttest occurred after reading the passage. In the *Read-Test-Read* condition, a posttest occurred after the first reading of the text passage and before the second reading of that passage. In all conditions, a 5-minute distractor task preceded a criterial test.

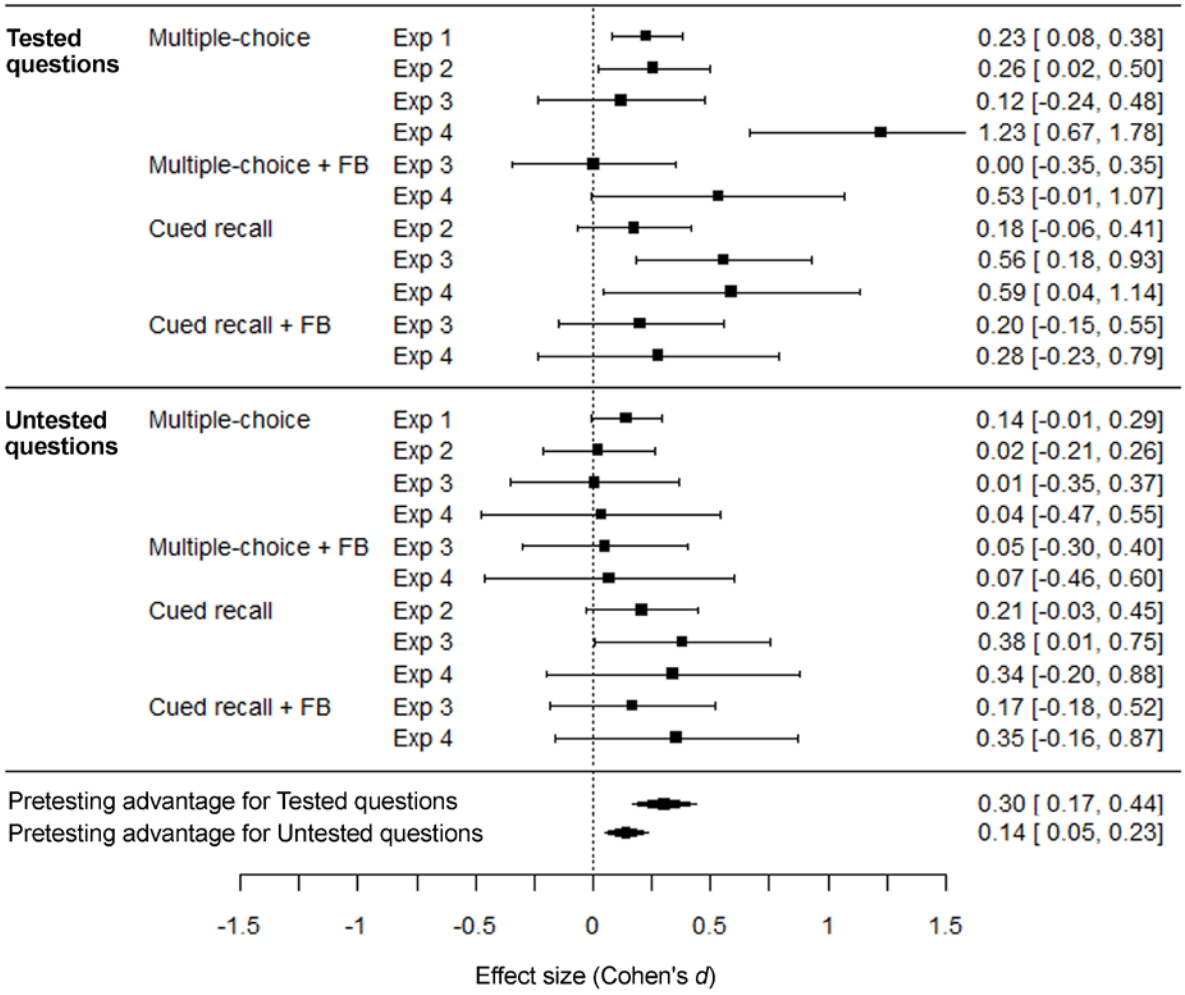

Figure 3

Forest plot of effect sizes (Cohen's *d*) with 95% confidence intervals for the difference between pretesting and posttesting on Tested and Untested criterial test questions, respectively, in Experiments 1-4. Positive *d* values indicate a pretesting advantage over posttesting. Exp = Experiment and FB = feedback.

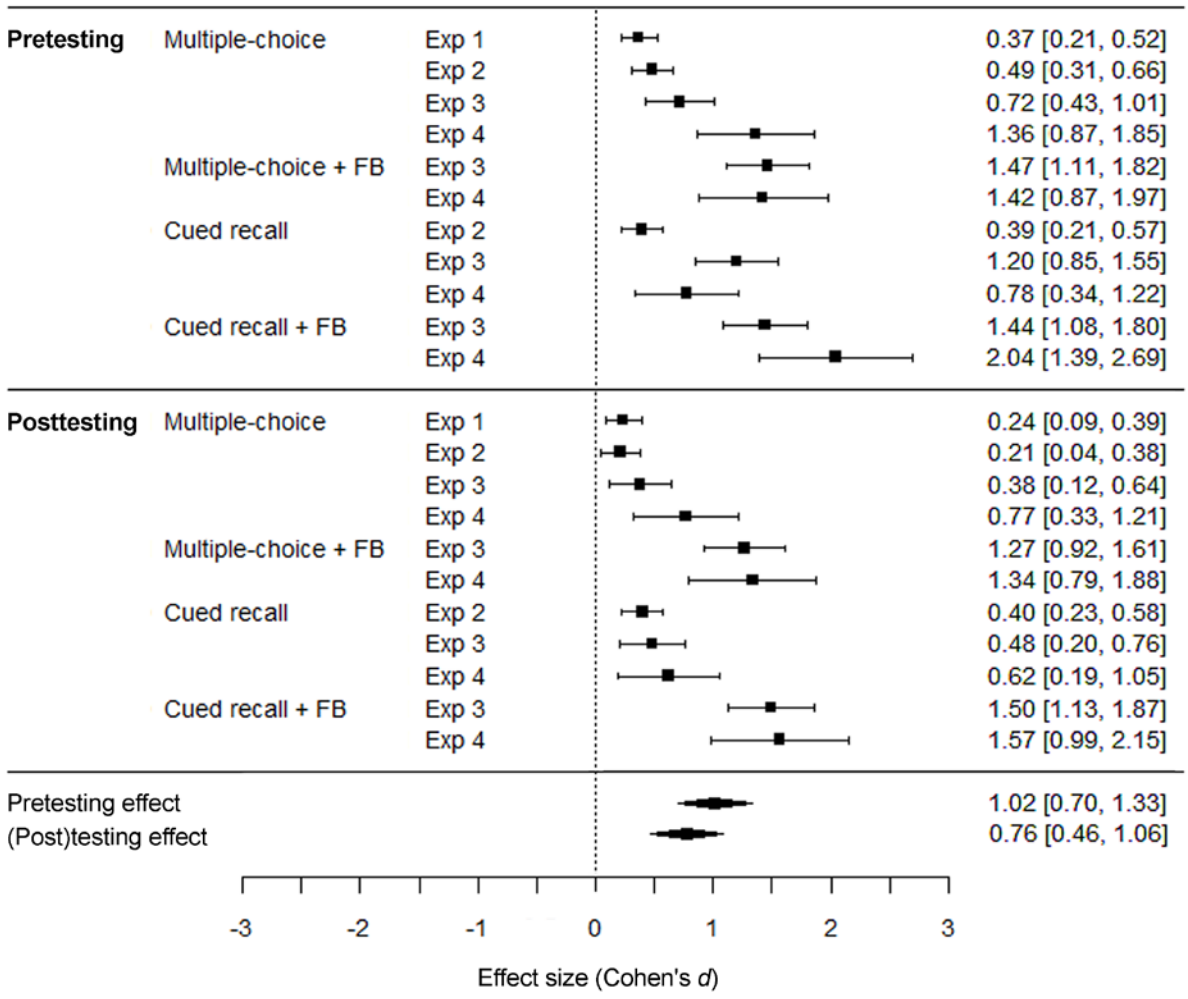

Figure 4

Forest plot of effect sizes (Cohen's *d*) with 95% confidence intervals for the pretesting and testing effects (i.e., performance on Tested versus Control questions) in Experiments 1-4. Positive *d* values indicating an advantage of pretesting or posttesting. Exp = Experiment and FB = feedback.

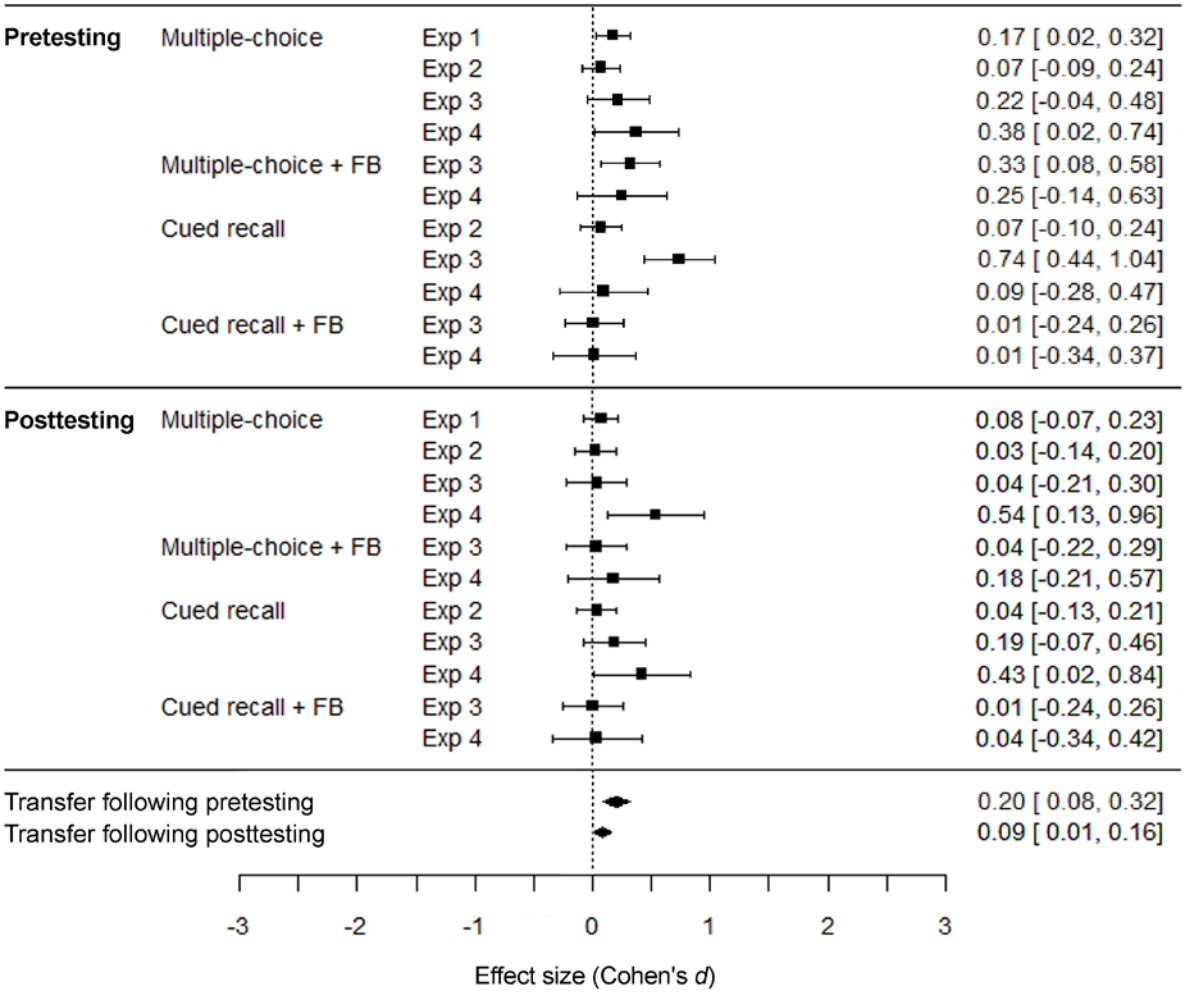

Figure 5

Forest plot of effect sizes (Cohen's *d*) with 95% confidence intervals for transfer effects following pretesting and posttesting (i.e., performance on Untested versus Control questions) in Experiments 1-4. Positive *d* values indicating an advantage of pretesting or posttesting. Exp = Experiment and FB = feedback.

Appendix A

Excerpts of the Saturn and Yellowstone National Park Text Passages (from Little, 2011)

Saturn

Saturn is the sixth planet from the Sun and the second largest planet in the Solar System. The planet is most well known for its beautiful system of planetary rings, which consist largely of water ice particles with smaller amounts of rocky debris and dust. Along with Jupiter, Uranus, and Neptune, Saturn is classified as a gas giant (also known as a Jovian planet, after the planet Jupiter).

The existence of Saturn has been known since prehistoric times: Saturn is the most distant planet that can be seen with the naked eye. Saturn gets its name from the Roman god Saturnus: the god of agriculture and harvest. The Romans considered Saturnus to be the equivalent of the Greek god, Kronos. Ancient Chinese cultures designated the planet Saturn as the 'earth star,' based upon the five elements which were traditionally used to classify natural elements. In Hindu astrology, Saturn is known as 'Sani' or "Shani'—the judge among all the planets.

Yellowstone National Park

Established in 1872, Yellowstone became America's first national park. The park is located at the headwaters of the Yellowstone River, for which it takes its name. In the eighteenth century, French trappers named the river "Roche Jaune" which is probably a translation of the Minnetaree name for "Rock Yellow River." Approximately 96% of the land area of Yellowstone National Park is located in the state of Wyoming, but the park extends into neighboring states of Idaho and Montana. Yellowstone is widely known for its wildlife and geothermal features: the park, itself, contains half of the world's geothermal features.

Evidence suggests that Aboriginal peoples have lived in the Yellowstone region for at least 11,000 years. The region is home to several Native American tribes including the Nez Perce, Crow, and Shoshone. European explorers first entered the region in the early nineteenth century. In 1806, John Colter left the Lewis and Clark Expedition to explore the region with a group of fur trappers. Upon seeing Yellowstone, he described it as a place of "fire and brimstone" due to the boiling mud, steaming rivers, and petrified trees.

Appendix B

Example Questions for the Saturn and Yellowstone National Park Text Passages (from Little, 2011)

Passage	Set A	Set B
Saturn	1. What planet lacks an internal magnetic field? a. **Venus** b. Earth c. Mercury d. Jupiter	1. On what planet is a day longer than a year? a. Venus b. Earth c. **Mercury** d. Jupiter
	2. Saturn's rings were first observed in what year? (at the time, however, they were not known to be rings). a. **1610** b. 1675 c. 1789 d. 1859	2. In what year did William Herschel discover Mimas and Enceladus, two moons of Saturn? a. 1610 b. 1675 c. **1789** d. 1859
	3. In 1655, who became the first scientist to suggest that Saturn is surrounded by a ring? a. Galileo b. Maxwell c. **Huygens** d. Keeler	3. Who first proposed that Saturn's rings aren't solid, but must instead be composed of many small particles? a. Galileo b. **Maxwell** c. Huygens d. Keeler
	4. The atmosphere of Saturn's rings is primarily composed of what element? a. **Oxygen** b. Hydrogen c. Helium d. Carbon	4. The body of Saturn is primarily composed of what element? a. Oxygen b. **Hydrogen** c. Helium d. Carbon
	5. Saturn was first visited in September of 1979 by which space probe? a. Voyager 1 b. Cassini-Huygens c. **Pioneer 11** d. Voyager 2	5. Which space probe collected data demonstrating wind speeds on Saturn exceeding 1,800 km/hour? a. **Voyager 1** b. Cassini-Huygens c. Pioneer 11 d. Voyager

(*appendix continues*)

Passage	Set A	Set B
Yellowstone National Park	1. What explorer left the Lewis and Clark Expedition to explore the region with a group of fur trappers? a. Raynolds b. **Colter** c. Bridger d. Hayden	1. What mountain man reported observing boiling springs, sprouting water, and a mountain of yellow rock, but was large ignored due to a reputation of being a 'spinner of yarns?" a. Raynolds b. Colter c. **Bridger** d. Hayden
	2. About 600 of what threatened species live within the Greater Yellowstone Ecosystem? a. elk b. bison c. **grizzly bears** d. grey wolf	2. What species makes up the largest population of a large mammal species in Yellowstone National Park? a. **elk** b. bison c. grizzly bears d. grey wolf
	3. Attacks by what tribe caused Colter to leave the Yellowstone region? a. Minnetaree b. Shoeshone c. **Blackfeet** d. Nez Perce	3. French trappers named Yellowstone River "Roche Jaune," probably a translation of what Native American tribe's name for Yellow Rock River? a. **Minnetaree** b. Shoeshone c. Blackfeet d. Nez Perce
	4. What is the tallest geyser in Yellowstone National Park? a. Old Faithful b. **Steamboat Geyser** c. Castle Geyser d. Daisy Geyser	4. What geyser is thought to be the oldest in the world? a. Old Faithful b. Steamboat Geyser c. **Castle Geyser** d. Daisy Geyser
	5. The majority of Yellowstone National Park resides in what state? a. Idaho b. South Dakota c. **Wyoming** d. Montana	5. The Black Hills region is found primarily in what state? a. Idaho b. **South Dakota** c. Wyoming d. Montana

Note. Cued recall versions of these questions were constructed by simply deleting the answer options and adding a response box for text entry. Boldface indicates correct answers. Materials adapted from Little (2011).