Educational Psychology Review (2023) 35:86
https://doi.org/10.1007/s10648-023-09803-8

INTERVENTION STUDY

Effortful Tests and Repeated Metacognitive Judgments Enhance Future Learning

Sara D. Davis[1] · Jason C. K. Chan[2]

Accepted: 3 August 2023 / Published online: 22 August 2023

Abstract

Prior testing can facilitate subsequent learning, a phenomenon termed the forward testing effect (FTE). We examined a metacognitive account of this effect, which proposes that the FTE occurs because retrieval leads to strategy optimizations during later learning. One prediction of this account is that tests that require less retrieval effort (e.g., multiple-choice relative to cued-recall) should lead to a smaller benefit on new learning. We examined the impact of interpolated multiple-choice or cued-recall testing (relative to no prior testing) on new learning of a four-section STEM text passage. The effect sizes associated with the FTE were numerically, though not significantly larger when the prior tests were cued-recall than multiple-choice, but only when interpolated judgments of learning were *not* queried. Further, when multiple-choice tests were made more difficult through lure similarity, the FTE was similarly increased. Finally, the FTE was eliminated entirely when participants provided four JOLs after reading each text section. We believe this elimination of the FTE stemmed from an increase in performance for the control participants induced by reactivity from repeated metacognitive queries requiring deep metacognitive reflection. Taken together, these experiments support a metacognitive account of FTE and have important implications for how educators and students should employ retrieval practice and leverage the benefits of metacognitive reflection to improve new learning.

Keywords Testing effect · Forward testing effect · Retrieval · Metacognition · JOL
Reactivity

All materials as well as data used for analyses can be found at the Open Science Framework at https://osf.io/ezjuv/.

✉ Sara D. Davis
sara.davis@unf.edu

[1] Department of Psychology, University of North Florida, Building 51, Room 3404, Jacksonville, FL 32224, USA

[2] Iowa State University, Ames, IA, USA

As researchers who study memory, cognitive psychologists are well positioned to identify techniques that can enhance learning in classroom contexts. A great number of techniques can enhance learning across domains**.** However, one of the most powerful tools yet identified to maximize learning is retrieval practice (see Adesope et al., 2017; McDermott, 2021, and Rowland, 2014, for reviews). Practicing retrieval on a previously-learned set of materials increases the likelihood that those tested materials will be remembered later. Further, retrieval can also enhance learning of material that is encoded *after* retrieval practice, a phenomenon called the forward testing effect (FTE, as compared to the general term, "testing effect", which typically describes the memorial benefits of testing on the materials encoded *before* retrieval practice; Pastotter et al., 2013; Szpunar et al., 2008; Pastötter & Bäuml, 2014). The FTE is robust, but its theoretical mechanisms are not yet well-understood. The purpose of the present study was to evaluate how metacognitive knowledge provided by retrieval practice guides new learning in favor of more effective study strategies. An important assumption of this metacognitive framework is that metacognition is tied to the act of retrieval, which makes easier tests such as recognition less likely to optimize strategy use than more difficult tests such as recall. To that end, we evaluated the ability of easier multiple-choice tests compared to more difficult cued-recall tests to benefit future learning.

Theoretical Mechanisms

In the typical paradigm, learners study several sections of material**.** Following each learning episode, learners engage in either interpolated tests or not. When tested, participants practice retrieval for some or all of the material in the prior learning opportunity. In the no-test comparison conditions, activities can include reviewing the previous material (i.e., restudy), completing a filler task (i.e., no test, interpolated break), or simply moving on to the next task (no test, no interpolated break). Importantly, all participants are tested on what they have studied during the final learning opportunity, and performance on this test allows researchers to evaluate the effect of prior activity on new learning. Learners who reviewed the earlier material with interpolated tests generally demonstrate much better recall of the final section materials (also called the criterial section) and reduced intrusions from the earlier sections (for a review, see Chan et al., 2018b)**.**

Chan et al. (2018b) classified theoretical accounts of this effect into four general categories (see also Pastötter & Bäuml, 2014; Yang et al., 2018 for other classifications): contextual segregation, attention, integration, and metacognitive accounts. The first of these, contextual segregation (Szpunar et al., 2008; see also Abel & Bäuml, 2016), suggests that interpolated tests help isolate the retrieved items from those studied in the criterial set based on a switch in task demand between encoding and retrieval (Davis et al., 2017; Kliegl & Bäuml, 2021; but see also Ahn & Chan, 2022). In contrast, attention accounts (see Pastötter et al., 2011, 2018) propose that testing helps learners sustain attention when they encode new information. Integration accounts propose that testing, by nature of strengthening the representation of previously studied material, aids new learning by facilitating the scaffolding of new material onto previous learning (Finn & Roediger, 2013; Jing et al., 2016, see Finn, 2017 for a review). It is important to note here that FTE is likely a multifaceted effect, and that the varying proposed mechanisms in the literature are not mutually exclusive. In fact, it is becoming increasing clear that a combination of several mechanisms give rise to enhanced new learning after testing (Chan et al. 2022; Kliegl & Bäuml, 2021), or that different materials or procedures can alter the extent to which each mechanism contributes (Ahn & Chan, 2022, in press).

Metacognitive Account

The metacognitive account proposes that learners may use their prior experience with tests to guide subsequent behavior, leading to enhanced new learning (Choi & Lee, 2020, Mazzoni et al., 1990; Sahakyan et al., 2004). These behavior changes can manifest as higher study time, which could be measured quantitatively, or as more qualitative strategy changes such as relating material to oneself, rereading, subvocalizing, etc. (see Cho & Powers, 2019, for a discussion of this distinction). In either case, repeated testing prompts learners to either repeat their previous encoding strategies throughout or to switch to more effective strategies as the experiment progresses (Panadero et al., 2017). In contrast, participants who are not tested may abandon the encoding strategies they used at the beginning of the encoding session (due to boredom, insufficient motivation, etc.), or they may be less likely to change strategies because the absence of testing reinforces less effortful behavior. As a result, the FTE could be attributed to an improvement by the tested participants, a deterioration by the non-tested participants, or both.

Some empirical findings support this account. For example, Chan et al., (2018a, 2020) and Yang et al., (2022) reported improved retrieval strategies (i.e., semantic and temporal clustering) for sections that follow retrieval practice. Cho and Powers (2019) reported that participants who were repeatedly tested tended to self-report more advantageous strategy use in the context of a general retrieval practice paradigm, although this finding failed to replicate in a recent study using an FTE paradigm (Ahn & Chan, in press). Thus, whereas there is some demonstrable evidence that prior testing can encourage strategy changes at retrieval (i.e., those based on clustering), the data on encoding strategies (i.e., those based on self-reports) is more equivocal.

Despite the importance of the metacognitive account to the understanding of the FTE, surprisingly little research has examined how repeated testing influences explicit metacognitive judgments in the context of the FTE. Yang et al. (2017) evaluated the FTE across several kinds of materials. Importantly, they directly measured aggregate judgments-of-learning (JOLs) after each list and measured self-paced study time of the material. Non-tested participants spent less time encoding across lists, but tested participants remained stable across lists. Interestingly, JOLs for the non-tested participants also declined across lists, whereas they remained steady for the tested participants. Szpunar et al. (2014) also asked for participants' judgments of learning after the encoding of a video lecture. Here, the JOLs were roughly equivalent for participants who had received interpolated tests during the lecture and those who had not. However, correspondence between predicted and actual scores was higher for participants who took interpolated tests than those who did not. The reason for the discrepancy between these two studies is unclear, and one goal of the present study is to further clarify the impact of testing on JOLs using educationally-relevant materials (see Rawson, 2015; for a discussion of the standard testing effect in complex materials)[1].

Multiple-Choice Testing and the FTE

According to the metacognitive account, learners should be more likely to alter their subsequent learning strategy if they are less successful during retrieval in the interpolated tests. One way to vary the likelihood of retrieval success is by varying the test difficulty. In the present experiments, we sought to manipulate difficulty by having participants complete interpolated recall vs. interpolated multiple choice. The types of tests used by researchers in this field have been relatively homogenous and, to our knowledge, no research has yet compared interim recognition to recall. In one exception, Yang et al., (2019) varied the type of interim tests participants completed in the test condition (e.g., recognition, classification, cued-recall) as well as the type of studied content (e.g., face-name pairs, Swahili-English word pairs, paintings and their artists) in each learning section. While they found that the FTE was robust even when test format and learning content changed (i.e., a transfer effect), they did not experimentally manipulate test format to evaluate whether one test format produced better new learning outcomes than others (relative to a nontested condition). Thus, this remains an important question for the present study to address.

In practice, multiple-choice tests are pervasive for classroom assessments, particularly for distance or remote courses (Gierl et al., 2017). However, multiple-choice testing may lead students to adopt less beneficial study strategies because they are easier than tests of cued-recall (i.e., short answers or fill-in-the-blank; see Roediger & Marsh, 2005). Students often overestimate their ability to remember information, and this over-confidence can lead to shorter self-paced study time (relative to students who are less over-confident) as well as the use of less effective study habits (Dunlosky & Rawson, 2012; Dunlosky et al., 2013; Kornell & Metcalfe, 2006, Smith & Karpicke, 2014. If students incorrectly assume that the material is well-learned because they took an easier interpolated test, they might devote less study time or effort to learning new information afterwards (Thiede & Dunlosky, 1994). In the case of the forward testing effect, this means that any benefit of testing (compared to a no prior test control) might be reduced when the interpolated tests are multiple-choice rather than recall.

[1] Note that Kubik et al. (2022) also examined JOLs in an FTE paradigm. However, their method employed item-by-item JOLs, which can recruit covert retrieval processes that mimic overt retrieval. In the case of the present study, we were interested primarily in aggregate judgments, which are less likely to elicit covert retrieval, particularly for complex material.

Overview of Experiments

The goal of the present experiments was to evaluate how metacognitive mechanisms contribute to the forward testing effect. To this end, we employed two procedures meant to test how different review types influence future learning as well as metacognitive knowledge. In four experiments, participants studied four sections of a passage about a scientific topic and took either a cued-recall test, a multiple-choice test, or no test for the first three sections. Participants were always tested on the fourth, criterial section, which allowed us to determine the impact of prior review type on the forward testing effect. According to the metacognitive account of FTE, participants who receive interpolated multiple-choice tests may become overconfident in their learning. Therefore, they might exert less effort or be less likely to optimize their subsequent learning strategies relative to participants who received the more difficult interpolated recall tests. Consequently, we predicted a smaller FTE when the interpolated test was multiple-choice relative to recall.

We also manipulated the format of the criterial test in addition to the format of the interpolated tests to examine how interpolated testing might affect students' performance on either a cued-recall criterial test or a multiple-choice criterial test. In the transfer-appropriate processing literature, performance can be facilitated when the conditions at encoding and retrieval match relative to when there is a mismatch (Morris et al., 1977). However, research from the testing effect literature supports a benefit of practicing effortful retrieval, regardless of the final test format. That is, learners typically benefit from initial tests that are more difficult, regardless of the difficulty/format of the final test (e.g., Carpenter & DeLosh, 2006; McDaniel et al., 2007). Given the variety of test formats used by learners and instructors in the classroom, the interplay between practice test format and final test format has important applied and theoretical implications.

We predicted that learners would devote more time to encode material in the presence of testing, and thus we allowed learners to read at their own pace so that we could measure self-regulated reading time. It is important to note here that self-regulated reading time is but one of many potential strategies that learners could employ or change in response to retrieval practice. However, other study strategies often require explicit reflection from participants (e.g., Ahn & Chan, in press), or may be otherwise difficult to quantitatively measure. While there are many reasons why learners may adjust their study time, we believe that this provides a relatively noninvasive procedure that can reflect shifts in strategy use based on metacognitive knowledge gained during previous tests.

We also asked participants to provide judgments of learning (JOLs), in which they indicated how likely they were to remember the material that they had just studied (King et al., 1980). As we noted earlier, Szpunar et al. (2014) and Yang et al. (2017) collected JOLs and found that interpolated testing either benefitted or did not affect metacognitive calibration, respectively. Therefore, we identified JOLs as an important measure of interest, given that a critical assumption of the metacognitive framework is that interpolated testing should aid students in more closely aligning their metacognitive judgments with their actual performance.

Lastly, it is important to note that providing JOLs is not always a neutral event. Instead, making JOLs sometimes alters subsequent learning behavior outside of the effects of retrieval practice (e.g., Mitchum et al., 2016; Zhao et al., 2021; see Rhodes, 2016 for a review). Specifically, providing evaluative judgments may induce reactivity, in which the prompt to reflect on future performance can result in participants' adjusting their encoding strategies on later learning opportunities. Given this possibility, we conducted Experiments 1a and 2a without requiring JOLs and Experiments 1b and 2b with JOLs. Experiments 1a and 1b are otherwise identical, as are Experiments 2a and 2b.

Experiment 1

All materials and data are available online at https://osf.io/ezjuv/.

Method

Participants

We determined minimum sample sizes for all experiments based on the effect size (Hedge's $g = 0.70$) reported in a meta-analysis by Chan et al. (2018b)[2]. G*Power software (Faul et al., 2007) indicated that 34 participants per between-subjects condition would be necessary to detect a single difference between two means with an alpha value of 0.05 and power of 0.80. Any deviations from the desired sample size occurred due to the randomization algorithm used by Qualtrics to route participants to each condition. In Experiment 1a, two hundred and thirty in-lab and online university participants (see Table 1 for the number of participants in each condition and experiment setting for each experiment) received partial course credit for their participation. We included the online participants to facilitate data collection, but the participants came from the same population of students, and sample did not influence criterial test performance[3], nor did it interact with any other variables influencing criterial test performance. Twenty-three participants were excluded from analysis (see Table 2), yielding a final sample size of 207. Experiment 1b included two-hundred and twenty-two participants, with 204 participants remaining after exclusions. Demographic characteristics for the final samples can be found in Table 3.

[2] The experiments reported here were not publicly pre-registered, but the methodology was presented in a dissertation proposal by S.D. to her doctoral program committee members, which include J.C.K. Importantly, the dissertation process served a similar function to a pre-registration, because S.D. was expected to adhere to the proposed methodology. Experiments 1a, 1b, and 2b were included in this proposal, and Experiment 2a was designed using the same sample size targets to ensure consistency between experiments.

[3] We conducted a 2 (Sample: Online vs. Lab) × 4 (Experiment) × 3 (Interpolated Test Condition: Cued-Recall, Easy Multiple-Choice, or No Test) × 2 (Criterial Test Condition: Cued-Recall vs. Multiple Choice) to determine the impact of sample on the primary dependent measure. The main effect of sample was not significant, nor were any interactions involving the Sample variable, F's < 1.87.

Table 1 Number of participants in each between-subjects condition in experiments 1 and 2

	Experiment			
Condition (PT-CT)	1a	1b	2a	2b
CR-CR	31 (5)	34 (10)	37 (11)	33 (10)
CR-MC	34 (9)	33 (10)	36 (11)	34 (10)
MC-CR	36 (12)	32 (10)	38 (13)	33 (10)
MC-MC	34 (8)	35 (10)	38 (10)	37 (10)
DMC-CR	–	–	45 (16)	39 (14)
DMC-MC	–	–	37 (11)	35 (11)
NT-CR	35 (11)	36 (9)	39 (11)	35 (11)
NT-MC	37 (10)	34 (10)	38 (12)	36 (12)

Number of online subjects in each condition appears in parentheses. PT indicates prior test condition and CT indicates criterial test condition. CR indicates cued-recall, MC indicates (easy) multiple-choice, DMC indicates difficult multiple-choice, and NT indicates no test

Table 2 Number of participants excluded and basis for exclusion in experiments 1 and 2

	Experiment			
Basis for Exclusion	1a	1b	2a	2b
Did not complete experiment	12	5	1	24
English not first language	8	11	10	15
Reported previously reading passage	1	–	8	4
Left experiment > 10 min (online only)	2	–	–	4

Design

Experiment 1 used a 2 (Criterial Test Type: Cued-Recall or Multiple-Choice) × 3 (Prior Review Type: Cued-Recall, Multiple-Choice, or No-Test) between-subjects design. The primary dependent variable was correct recall or recognition on the criterial test. This general design was identical between Experiments 1a and 1b, with the only procedural difference between them being the inclusion of aggregate JOL judgments provided in Experiment 1b as described below.

Materials and Procedure

A prose passage about the development, use, and components of lasers (11.3 Flesch-Kinkaid grade level) from an online textbook website written for the general public (Woodford, 2021) was divided into four sections ($M_{\text{Section1}}=753$ words, $M_{\text{Section2}}=754$ words, $M_{\text{Section3}}=748$ words, $M_{\text{Section4}}=750$ words). Later sections built upon the earlier ones, such that learning prior sections should facilitate learning of the subsequent ones. Although this structure prevented us from counterbalancing the materials across sections (e.g., the information covered in the fourth

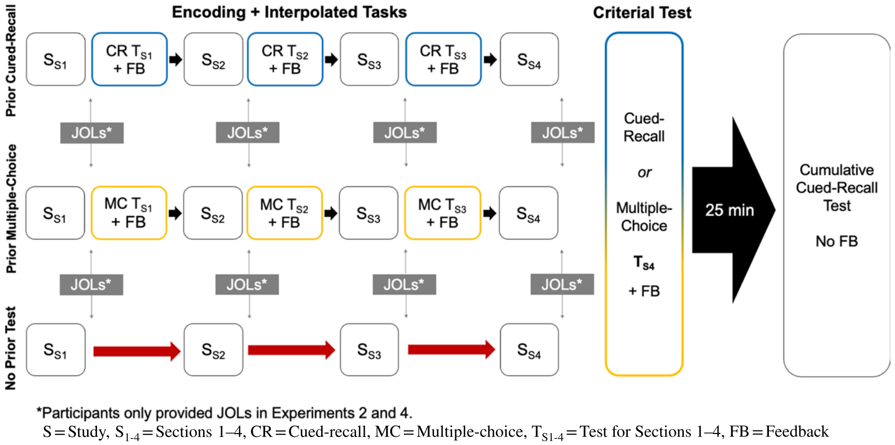
Table 3 Self-identified sample demographic characteristics in experiments 1 and 2

	Experiment			
	1a	1b	2a	2b
Age	19 years	19 years	19 years	19 years
Gender				
Female	54%	56%	54%	51%
Male	44%	38%	44%	48%
Declined to Identify	2%	5%	< 1%	1%
Race				
White	83%	77%	81%	81%
Black	5%	4%	5%	3%
Asian or Asian-American	4%	6%	6%	5%
Hispanic or Latinx	2%	1%	4%	4%
Mixed-Race	1%	2%	< 1%	5%
Pacific Islander	< 1%	–	–	–
Native American	–	< 1%	< 1%	–
Other Race	–	–	1%	< 1%
Declined to Identify	4%	6%	2%	1%

Gender/sex was an open-ended question that was coded as Male, Female, Other, or Decline to Identify. Race was a multiple-choice question, and participants could select more than one race. A response to the demographic questions was not required

section could not be presented in a different section), it was ecologically realistic and did not affect our ability to make comparisons across the review conditions.

A graphical depiction of the procedure is presented in Fig. 1. Participants received instructions that they would be reading a series of text passages at their own pace, and they should learn them as if they were reading a textbook in one of their courses. They were also told that the computer would randomly determine how they would



*Participants only provided JOLs in Experiments 2 and 4.
S = Study, S$_{1\text{-}4}$ = Sections 1–4, CR = Cued-recall, MC = Multiple-choice, T$_{S1\text{-}4}$ = Test for Sections 1–4, FB = Feedback

Fig. 1 A graphical depiction of the method used in experiments 1 and 2

review the material after each text section: with a short-answer (i.e., cued-recall) test, a multiple-choice test, or no test. In actuality, participants performed the same review activity for the first three sections (depending on the assigned condition) and always took either a multiple-choice test or a short-answer test for the fourth section. All participants were told that they would take a final cumulative test.

Eight open-ended test questions were designed for each section. For the cued-recall tests, participants were prompted to provide a short response to the question (e.g., What part of the laser is stimulated by the flash tube? _______, in which the participant should answer "gain medium.") In contrast, the multiple-choice tests presented the correct answer along with three foils in a random order (e.g., a. gain medium, b. event potential, c. sublimator, d. actuator), and participants were instructed to select the correct answer with a mouse click. Critically, the question stem was identical for both the multiple-choice and cued-recall questions. During all tests, each question was followed by the correct answer as feedback regardless of test format. Encoding of these feedback trials was self-paced, and subjects used the mouse to advance to the next question. Participants in the no prior test condition simply advanced to the next section without completing an interpolated task.

Following completion of the criterial test for the final section, participants completed two tasks (i.e., alpha span; Craik, 1986; and backward digit span; Hayslip & Kennelly, 1982) as distractors for 15 min before the final test. They then received the final cumulative cued-recall test, which contained both repeated questions from the interpolated tests and new questions. This cumulative test was self-paced, and questions were presented in a random order. Finally, all participants answered a short demographics questionnaire.

The procedure for Experiment 1b was identical to Experiment 1a except that all participants were prompted to provide four aggregate JOLs (i.e., estimate their future performance) after reading each text section. The four JOL questions asked participants to predict their performance on an immediate cued-recall test, an immediate multiple-choice test, a final cumulative cued-recall test, and a final cumulative multiple-choice test by using a sliding scale from 0–100[4]. The time to make each judgment was self-paced but was not recorded. The four JOL probes appeared one at a time and in a random order for each section.

Results

For each analysis, we report data for both null hypothesis significance testing (NHST) and Bayes' Factors (BF). We report BF_{10} when the NHST result was significant at two-tailed alpha = 0.05 and BF_{01} when the NHST result was not significant, so that a larger BF always indicated more support for the reported effect. All Bayesian analyses were conducted using the default prior parameters in JASP (JASP

[4] articipants were never given a delayed multiple-choice test in actuality, but we included questions regarding a delayed cued-recall and multiple-choice test to be consistent with the questions for the immediate tests.

Table 4 Mean performance on the cumulative final test in experiments 1 and 2

	Experiment			
	1a	1b	2a	2b
Prior Cued-Recall	.39 (.16)	.42 (.17)	.46 (.17)	.43 (.20)
Prior Difficult Multiple-Choice	–	–	.43 (.16)	.41 (.17)
Prior Easy Multiple-Choice	.35 (.20)	.36 (.16)	.39 (.15)	.43 (.18)
No Prior Tests	.22 (.15)	.29 (.14)	.25 (.14)	.29 (.12)

Standard deviations are in parentheses

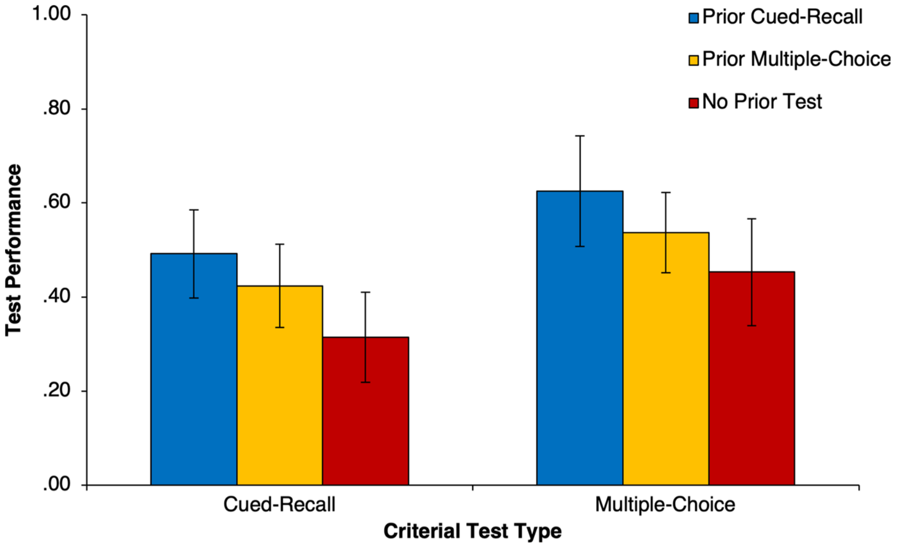
Team, 2023): a point-null vs. an H_1 featuring a zero-centered Cauchy distribution with a interquartile range of $r = \pm 0.707$.

We opted not to include comprehensive analyses of the cumulative test data for all experiments because the backward testing effect was not the focus of our study, and because the backward testing effect could be attributed to both retrieval practice and re-exposure during feedback. Data regarding the cumulative test can be found in Table 4, and as expected, there was a backward testing effect in all experiments. These data will not be discussed further, all Fs > 11.59, ps < 0.001, η_p^2s > 0.10, BF_{10}s > 1084.35.

Criterial Test Performance

Experiment 1a

A 2 (Criterial Test Type: Cued-Recall or Multiple-Choice) × 3 (Prior Review Type: Cued-Recall, Multiple-Choice, or No Prior Test) between-subjects Analysis of Variance (ANOVA) examined whether the forward testing effect was observed on the criterial test (see Fig. 2). There was a main effect of Criterial Test Type, such that performance was higher overall when the criterial test was multiple-choice ($M = 0.54$) than when it was cued-recall ($M = 0.41$), $F(1, 201) = 10.42$, $p = 0.001$, $\eta_p^2 = 0.05$, $BF_{10} = 19.81$, confirming that our multiple-choice test was easier than our cued-recall test. There was also a main effect of Prior Review Type, $F(2, 201) = 6.47$, $p = 0.002$, $\eta_p^2 = 0.06$, $BF_{10} = 15.04$. The interaction between Criterial Test Type and Prior Review Type was not significant, $F(2, 201) = 0.04$, $p = 0.96$, $BF_{01} = 10.62$. Post-hoc tests revealed a forward testing effect when comparing the prior cued-recall condition ($M = 0.56$) to the no prior test condition ($M = 0.39$), $t(135) = 3.42$, $p = 0.001$, $d = 0.59$, $BF_{10} = 32.91$. There was also a marginal forward testing effect when comparing prior multiple-choice ($M = 0.48$) to no prior test, $t(140) = 1.95$, $p = 0.053$, $d = 0.33$, $BF_{10} = 1.02$, although this effect was approximately half that of prior cued-recall. Lastly, participants in the prior cued-recall condition numerically outperformed those in the prior multiple-choice condition, but the effect did not reach conventional standards of significance, $t(133) = 1.69$, $p = 0.094$, $d = 0.29$, $BF_{01} = 1.49$. Note, however, that the FTE is typically represented as a comparison



Bars represent descriptive 95% confidence intervals

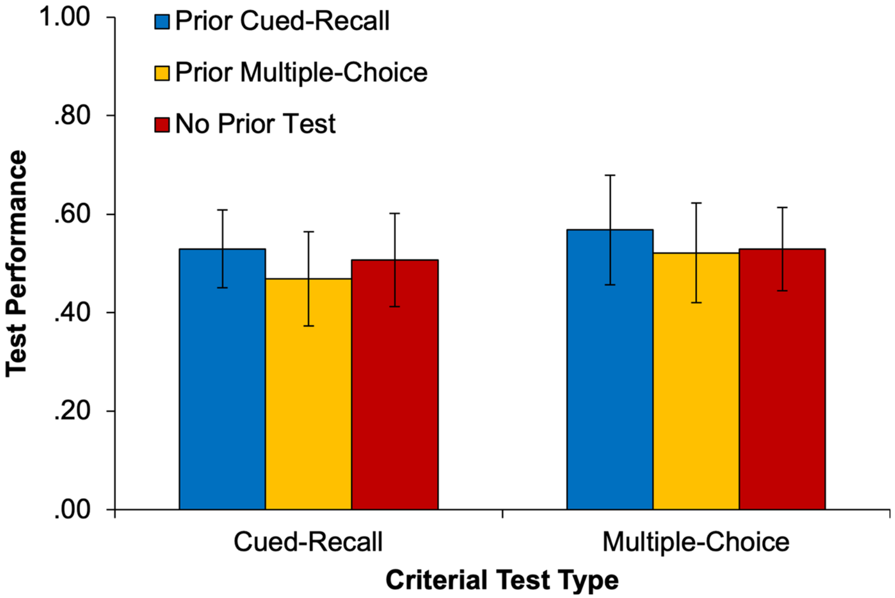
Fig. 2 Criterial test performance as a function of prior review type in experiment 1a

between a tested condition and a non-tested condition (Chan et al., 2018b). Given that the magnitude of the effect between the two tested conditions was smaller than this usual comparison, this nonsignificant effect is unsurprising.

Taken together, these findings are consistent with the metacognitive account and suggest that cued-recall interpolated tests is more effective at improving future learning than multiple-choice interpolated tests. Further, interpolated testing promoted subsequent learning regardless of whether the criterial test was recall or multiple-choice.

Experiment 1b

A 2 (Criterial Test Type) × 3 (Prior Review Type) between-subjects ANOVA showed no main effect of Prior Review Type, $F(2, 198) = 0.62$, $p = 0.55$, $\eta_p^2 = 0.01$, $BF_{01} = 11.82$; no main effect of Criterial Test Type, $F(1, 198) = 0.92$, $p = 0.34$, $\eta_p^2 = 0.01$, $BF_{01} = 4.35$, and no interaction, $F(2, 198) = 0.05$, $p = 0.95$, $\eta_p^2 < 0.001$, $BF_{01} = 10.51$. As can be seen in Fig. 3, there was no forward testing effect in Experiment 1b ($M_{CR} = 0.55$, $M_{MC} = 0.50$, $M_{NT} = 0.52$), nor was there an effect of multiple-choice ($M = 0.50$) vs. cued-recall ($M = 0.54$) difficulty on the criterial test. We will discuss these surprising findings in more depth in the General Discussion, but to preview, we suspect that requiring participants to provide JOLs might have induced metacognitive introspection that triggered a strategy change even for participants in the no-test condition — i.e., JOL reactivity – thus eliminating the FTE.



Bars represent descriptive 95% confidence intervals

Fig. 3 Criterial test performance as a function of prior review type in experiment 1b

Reading Time Data

Experiment 1a

Before conducting an ANOVA on reading times, we first used the interquartile range (i.e., IQR) rule to identify and remove outliers, because very long or very short reading times could indicate noncompliance with the experimental procedure. The IQR rule is suitable for data with non-normal distributions (Dunn, 2021), which stipulates that data points that lie 1.5 * IQR below the first quartile and 1.5 * IQR over the third quartile might be outliers. We applied this rule to remove outliers per condition (i.e., Prior Review Type) and for each text section. For the repeated-measures ANOVA reported below, we retained 63 participants in the prior cued-recall condition (two were removed), 62 in the prior multiple-choice condition (nine were removed), and 67 participants in the no-test condition (five were removed).

A 3 (Prior Review Type) × 4 (Text Section) repeated measures ANOVA revealed a significant main effect of prior review type, $F(2, 188) = 8.75$, $p < 0.001$, $\eta_p^2 = 0.09$, $BF_{10} = 113.44$, a nonsignificant main effect of text section, $F(3, 564) = 1.99$, $p = 0.114$, $\eta_p^2 = 0.01$, $BF_{01} = 12.54$, and most importantly, an interaction, $F(3, 564) = 3.78$, $p = 0.001$, $\eta_p^2 = 0.04$, $BF_{10} = 11.78$. This interaction was driven by the fact that reading time remained relatively constant across text sections for participants in the prior cued-recall condition (211 s, 212 s, 219 s, 220 s), $F(3, 186) = 0.47$, $p = 0.706$, $\eta_p^2 < 0.01$, $BF_{01} = 29.57$, and the prior multiple-choice condition (217 s, 229 s, 230 s, 210 s), $F(3, 180) = 1.855$, $p = 0.139$, $\eta_p{}^2 = 0.03$, $BF_{01} = 5.05$, but reading time declined markedly across text sections for participants in the no prior test condition (185 s, 156 s, 156 s, 145 s), $F(3, 198) = 8.13$, $p < 0.001$, $\eta_p{}^2 = 0.11$, $BF_{10} = 437.28$. These results are consistent with the metacognitive account, which postulates that interpolated testing can promote later learning by inducing beneficial encoding behaviors. Criterial section reading time was also positively correlated with test performance, $r = 0.36$, $p < 0.001$, $BF_{10} = 105756.70$, and this positive relationship was observed in all conditions, $r_{\mathrm{CR}} = 0.38$, $r_{\mathrm{MC}} = 0.28$, $r_{\mathrm{NT}} = 0.30$, $p\mathrm{s} < 0.025$, $BF_{10}\mathrm{s} > 1.80$.

Experiment 1b

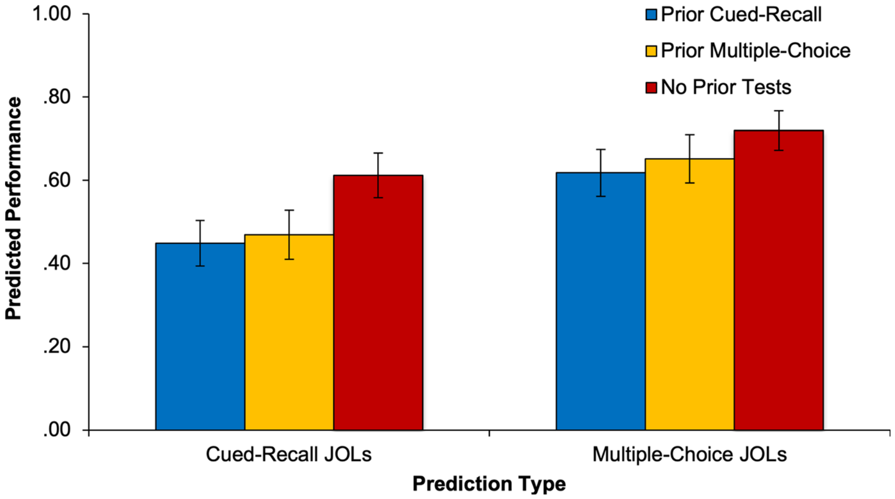
Outlying reading time data were removed using the same procedure as Experiment 1a. We retained 60 participants in the prior cued-recall condition (seven were removed), 58 in the prior multiple-choice condition (nine were removed), and 59 participants in the no-test condition (11 were removed).

Using the same ANOVA analysis as Experiment 1a, there were no significant main effects or interactions observed, all $F\mathrm{s} < 1.87$, $p\mathrm{s} > 0.133$, $BF_{01}\mathrm{s} > 6.97$. Most critically, reading time did not decline across text sections for participants regardless of prior review type, $F\mathrm{s} < 1.01$, $p\mathrm{s} > 0.390$, $BF_{01}\mathrm{s} > 13.69$. Thus, unlike Experiment 1a, test condition had no impact on reading times, and criterial section reading time was no longer correlated with test performance, $r = 0.09$, $p = 0.22$, $BF_{01} = 5.38$. In fact, a meaningful positive correlation was absent in all conditions, $r_{\mathrm{CR}} = -0.13$, $r_{\mathrm{MC}} = 0.05$, $r_{\mathrm{NT}} = 0.15$, $p\mathrm{s} > 0.234$, $BF_{01}\mathrm{s} > 3.32$.

Criterial Section Judgments of Learning

We next examined participants' JOL ratings for the criterial test in Experiment 1b. We took this approach because the criterial test JOLs could be compared to actual test performance for all participants, but note that other analysis approaches (e.g., by examining the average JOLs across all four sections of text) produce similar outcomes. A 3 (Prior Review Type) × 2 (Predicted Test Type) mixed ANOVA revealed a main effect of Predicted Test Type, $F(1, 201) = 167.05$, $p < 0.001$, $\eta_p{}^2 = 0.45$, $BF_{10} = 1.53 \text{ x } 10^{24}$, a main effect of Prior Review Type, $F(2, 201) = 7.58$, $p < 0.001$, $\eta_p{}^2 = 0.07$, $BF_{10} = 43.74$, and an interaction, $F(2, 201) = 3.84$, $p = 0.023$, $\eta_p{}^2 = 0.04$, $BF_{10} = 1.39$.

Inspection of Fig. 4 shows that the main effect of Predicted Test Type was driven by higher performance predictions for multiple-choice tests ($M = 0.66$) than cued-recall ($M = 0.51$). The main effect of prior review type was driven by lower JOLs given by participants who had taken cued-recall tests ($M = 0.53$) and multiple-choice tests ($M = 0.56$) relative to their no prior test counterparts ($M = 0.67$). The interaction appears to reflect that the retrieval-based reduction in JOLs was particularly potent when participants considered future cued-recall performance relative to future multiple-choice performance.



Bars represent descriptive 95% confidence intervals

Fig. 4 JOL's provided in experiment 1b for immediate tests as a function of prior review type and prediction type

Discussion

The findings from Experiment 1a and b provide support for the idea that testing influences subsequent strategy use (as evidenced by the reading time data) that improves memory. Prior cued-recall testing enhanced new learning *regardless of whether participants took a cued-recall or a multiple-choice criterial test*. Most importantly, although prior multiple-choice testing enhanced learning of the final section, the effect size was approximately half that of prior cued recall (but note that the direct comparison between these two test conditions was not significant).

The most important and surprising result here was that unlike Experiment 1a, there was no forward testing effect in Experiment 1b. Comparing the means for the criterial test between Experiments 1a and b, there was an increase in performance for participants in the no-test condition (M_{NT}=0.39 in Experiment 1a vs. $M_{\mathrm{NT}\,=}$ 0.52 in Experiment 1b) but not for participants in the tested conditions (M_{CR}=0.56, M_{MC}=0.48 in Experiment 1a vs. M_{CR}=0.55, M_{MC}=0.50 in Experiment 1b). Importantly, making judgments of learning can sometimes eliminate the benefits of testing by increasing performance of participants in control conditions, particularly when the materials are interrelated (e.g., Double et al., 2018; Dougherty et al., 2005, 2018; Janes et al., 2018; Myers et al., 2020; Zhao et al., 2021). Our results would be the first demonstration of this JOL reactivity effect in the context of the FTE with realistic, STEM-based learning material. It is possible that querying judgments of learning four times per section may have encouraged participants to reflect on their learning and change strategies, benefitting participants in the no-test condition the most. The finding that participants performed similarly regardless of criterial test format (M_{CR}=0.50, M_{MC}=0.54 in Experiment 2; M_{CR}=0.41, M_{MC}=0.57 in Experiment 1a) is more difficult to explain. We return to this idea in more detail in the General Discussion.

Despite the null effects of the review type manipulations on criterial test accuracy, interpolated testing was associated with lower metacognitive judgments. Individuals in the no prior test condition reported higher JOLs than those in the test conditions (as predicted by the metacognitive account). However, it is important to note that these differences in metacognitive judgments did not translate into actual performance differences for the criterial test.

Beyond JOL-induced reactivity, a question that remains is how one might make multiple-choice testing more effective at enhancing learning (or how one might design multiple-choice tests to result in better calibration), given their pervasive use in educational contexts. One way to do this is to reduce recognition fluency by increasing the difficulty of the tests. To this end, we increased the competitiveness of the lures on the multiple-choice tests in Experiment 2. We also sought to replicate Experiments 1a and 1b by implementing the new procedure both with (Experiment 2b) and without (Experiment 2a) the JOL probes included.

Experiment 2

Method

Participants, Design, Materials, and Procedure Participants in Experiment 2a were 327 university students, with 308 participants remaining in the final analysis. Three-hundred and twenty-nine participants took part in Experiment 2b, with 282 participants remaining in the final sample (see Tables 1–3).

The materials, design, and procedure were similar to Experiment 1, with the addition of a prior *difficult* multiple-choice condition for the interpolated tests. Thus, the design was a 2 (Criterial Test Type: Cued-Recall or Multiple-Choice) × 4 (Prior Review Type: Cued-Recall, Easy Multiple-Choice, Difficult Multiple-Choice, or No-Test) between-subjects design. Experiment 2a replicated Experiment 1a with no JOL questions between the sections of text, and Experiment 2b replicated Experiment 1b by including these questions.

The Easy Multiple-Choice condition in Experiment 2 was identical to the Multiple-Choice condition in Experiment 1. The lures from these multiple-choice questions were modified for the Difficult Multiple-Choice condition, so that they were more competitive with the correct answer, although the stems remained the same (see Little et al., 2012 for a description of this method). The criterial multiple-choice test was the same as in Experiment 1 (i.e., it contained the less-competitive lures).

Results

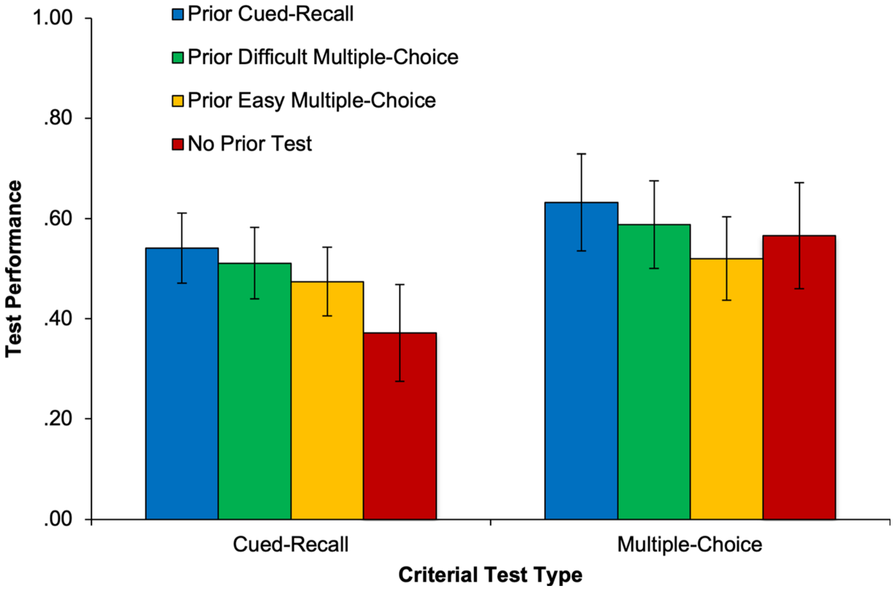
Manipulation Check

To verify that the manipulation of multiple-choice question difficulty was indeed successful, we examined aggregate interpolated test performance for Lists 1–3 in Experiment 2a. Of critical importance, the difference between the two multiple-choice conditions was significant, $t(156) = 2.84$, $p = 0.015$, $d = 0.45$, $BF_{10} = 7.41$, as was the difference between the cued-recall and difficult multiple choice conditions, $t(153) = 9.57$, $p < 0.001$, $d = 1.54$, $BF_{10} = 1.58 \times 10^{14}$.

Criterial Test Performance

Experiment 2a

We performed a 2 (Criterial Test Type) × 4 (Prior Review Type) between-subjects ANOVA (see Fig. 5). Similar to Experiment 1a, there was a main effect of Criterial Test Type, $F(1, 300) = 10.90$, $p = 0.001$, $\eta_p^2 = 0.04$, $BF_{10} = 18.74$, such that participants achieved higher performance on the multiple-choice criterial test ($M = 0.58$) than on the cued-recall criterial test ($M = 0.47$). There was again a main effect of Prior Review Type as well, $F(3, 300) = 2.85$, $p = 0.04$, $\eta_p^2 = 0.03$, $BF_{10} = 0.61$, with a significant FTE for the prior cued-recall condition ($M = 0.59$) relative to the no prior test condition ($M = 0.47$), $t(148) = 2.41$, $p = 0.02$, $d = 0.39$, $BF_{10} = 2.46$, a marginal FTE for the difficult multiple-choice condition ($M = 0.55$) relative to the no prior test condition, $t(157) = 1.66$, $p = 0.10$, $d = 0.26$, $BF_{10} = 0.61$, and a nonsignificant FTE for the easy multiple-choice condition ($M = 0.50$) relative to the no prior test condition, $t(151) = 0.62$, $p = 0.54$, $d = 0.10$, $BF_{01} = 4.81$. The interaction was not significant, $F(3, 300) = 1.08$, $p = 0.36$, $\eta_p^2 = 0.01$, $BF_{01} = 8.22$.



Bars represent descriptive 95% confidence intervals

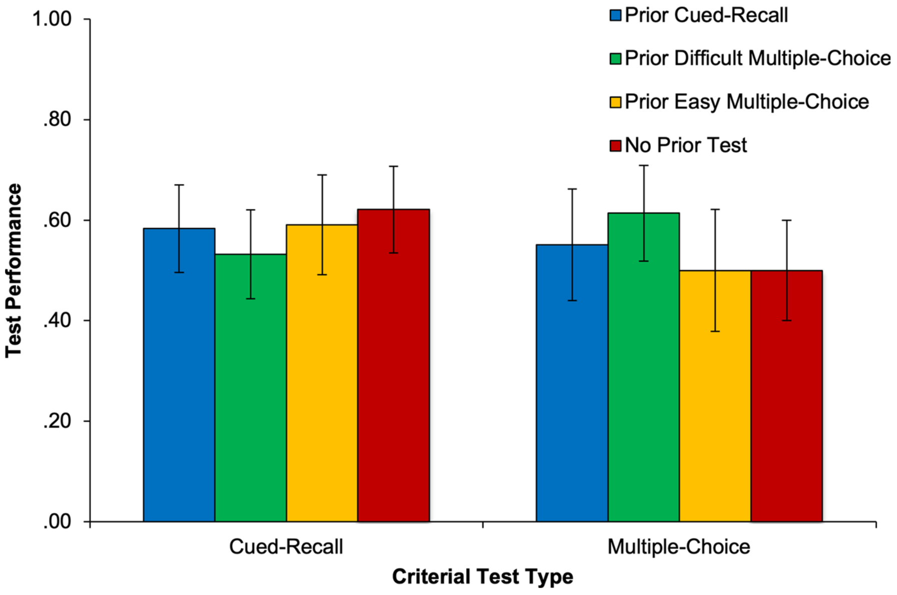
Fig. 5 Criterial test performance as a function of prior review type in experiment 2a

Thus, as predicted by the metacognitive account, prior cued-recall produced the largest benefit, but increasing the difficulty of the multiple-choice tests also improved performance relative to no testing.

We also examined whether participants in the prior cued-recall condition outperformed those in the two prior multiple-choice conditions, although recall that these experiments were not designed to detect small differences between the tested conditions. Here, prior cued-recall led to superior criterial test performance when compared to prior easy multiple-choice, $t(147) = 2.17$, $p = 0.03$, $d = 0.36$, $BF_{10} = 1.49$, but not when compared to prior difficult multiple-choice, $t(153) = 0.96$, $p = 0.34$, $d = 0.15$, $BF_{01} = 3.79$.

Experiment 2b

Importantly, we replicated the absence of the FTE in Experiment 2 (see Fig. 6). There was no main effect of Prior Review Type, $F(3, 274) = 0.11$, $p = 0.95$, $\eta_p^2 = 0.001$, $BF_{01} = 52.99$, no main effect of Criterial Test Type, $F(1, 274) = 0.26$, $p = 0.61$, $\eta_p^2 = 0.01$, $BF_{01} = 4.25$, nor was there an interaction, $F(3, 274) = 1.61$, $p = 0.19$, $\eta_p^2 = 0.02$, $BF_{01} = 3.95$. These data provide further evidence that the forward effect of testing was eliminated because participants were asked to make four JOLs after each study segment. Further, we replicated the surprising finding that the difference in criterial test format disappeared when participants were asked to make JOLs.



Bars represent descriptive 95% confidence intervals

Fig. 6 Criterial test performance as a function of prior review type in experiment 2b

Reading Time Data

Experiment 2a

After outliers were removed, we retained 65 participants in the prior cued-recall condition (eight were removed), 72 in the prior difficult multiple-choice condition (10 were removed), 73 in the prior easy multiple-choice condition (three were removed), and 74 participants in the no-test condition (three were removed). A 4 (Prior Review Type) × 4 (Text Section) repeated measures ANOVA showed a significant main effect of prior review type, $F(3, 280) = 4.84$, $p = 0.003$, $\eta_p^2 = 0.04$, $BF_{10} = 13.33$. The main effect of text section was just significant in NHST terms, but the *BF* favored the null, $F(3, 840) = 2.94$, $p = 0.032$, $\eta_p^2 < 0.01$, $BF_{01} = 6.45$. Moreover, unlike Experiment 1, the interaction was not significant, $F(3, 840) = 1.56$, $p = 0.181$, $\eta_p^2 = 0.03$, $BF_{01} = 6.84$.

To further scrutinize the data, we examined reading time across sections for participants in each condition. Similar to the data in Experiment 1, reading time remained relatively stable for participants in the prior cued-recall condition (207 s, 219 s, 216 s, 203 s), $F(3, 192) = 1.64$, $p = 0.706$, $\eta_p^2 < 0.01$, $BF_{01} = 29.57$, and the prior easy multiple-choice condition (195 s, 199 s, 203 s, 199 s), $F(3, 216) = 0.42$, $p = 0.739$, $\eta_p^2 < 0.01$, $BF_{01} = 37.32$. Further, reading time declined for participants in the no prior test condition (175 s, 158 s, 163 s, 156 s), $F(3, 219) = 3.57$, $p = 0.015$, $\eta_p^2 = 0.05$, $BF_{10} = 1.46$, although the effect was smaller than that observed in Experiment 1a. Somewhat unexpectedly, reading time also declined for participants in the prior difficult multiple-choice condition, although the effect was small and not significant (207 s, 205 s, 199 s, 190 s), $F(3, 213) = 2.30$, $p = 0.078$, $\eta_p^2 = 0.03$, $BF_{01} = 3.36$.

These data are generally consistent with our data from Experiment 1 as well as the assumptions under a metacognitive framework. There was again a significant positive correlation between criterial section reading times and test performance, $r = 0.31$, $p < 0.001$, $BF_{10} = 2.55$ x 10^6, and similar to Experiment 1a, a positive correlation was evident in every prior review condition, $r_{\mathrm{CR}} = 0.34$, $r_{\mathrm{DMC}} = 0.26$, $r_{\mathrm{EMC}} = 0.28$, $r_{\mathrm{NT}} = 0.32$, *p*s < 0.021, BF_{01}s > 1.95.

Experiment 2b

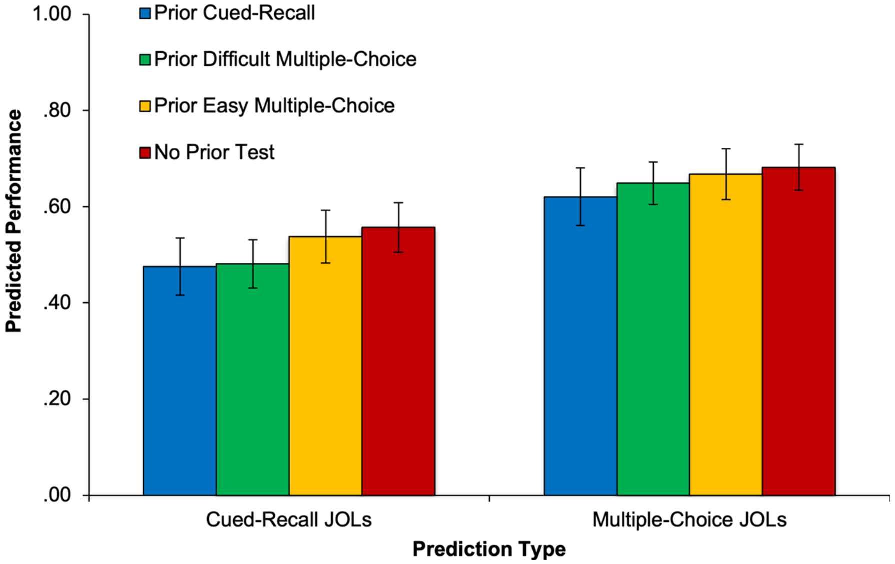
We retained 60 participants in the prior cued-recall condition (10 were removed), 58 in the prior difficult multiple-choice condition (seven were removed), 58 in the prior easy multiple-choice condition (nine were removed), and 59 participants in the no-test condition (five were removed).

A 4 (Prior Review Type) X 4 (Text Section) repeated measures ANOVA showed no interaction, $F(3, 741) = 0.45$, $p = 0.909$, $\eta_p^2 < 0.01$, $BF_{01} = 1227$. The main effect of text section was also not significant, $F(3, 741) = 2.33$, $p = 0.073$, $\eta_p^2 < 0.01$, $BF_{01} = 11.52$. Unexpectedly, there was a (just) significant main effect of prior review type, although the *BF* was completely neutral, $F(3, 247) = 2.73$, $p = 0.045$, $\eta_p^2 = 0.03$, $BF_{10} = 1.07$. When we examined the data for participants in each condition separately, it became clear that the data were largely consistent with those from Experiment 2, such that reading time did not significantly decline across text sections regardless of prior review type, $F\text{s} < 1.53$, $p\text{s} > 0.210$, $BF_{01}\text{s} > 7.59$.

Lastly, we examined whether criterial section reading time was predictive of test performance. Unlike Experiment 1b, in which reading time and test performance were uncorrelated, there was a moderate correlation here, $r = 0.25$, $p < 0.001$, $BF_{10} = 684.62$. When examining each prior review condition separately, participants in the prior cued-recall and no prior test conditions showed a significant positive correlation, $r\text{s} > 0.34$, $p\text{s} < 0.004$, $BF_{01}\text{s} > 9.50$, but those in the prior multiple-choice conditions did not, $r\text{s} < 0.21$, $p\text{s} < 0.089$, $BF_{01}\text{s} > 1.58$.

Criterial Section Judgments of Learning

A 4 (Prior Review Type) × 2 (Predicted Test Type) ANOVA on JOLs for the criterial test (see Fig. 7) revealed a main effect of predicted test type, $F(1, 278) = 240.27$, $p < 0.001$, $\eta_p^2 = 0.46$, $BF_{10} = 1.38 \text{ x } 10^{36}$, but neither the main effect of Prior Review Type nor the interaction were significant, $F\text{s} < 1.69$, $p\text{s} > 0.170$, $\eta_p^2\text{s} < 0.02$, $BF_{01}\text{s} > 1.55$. Consistent with the results in Experiment 1b, participants provided higher JOL estimates for multiple-choice tests ($M = 0.66$) than for cued-recall tests ($M = 0.51$). Unlike Experiment 1b, the main effect of Prior Review Type was not significant here, although the numerical pattern was similar.



Bars represent descriptive 95% confidence intervals

Fig. 7 JOL's Provided in experiment 2b for immediate tests as a function of prior review type and prediction type

Discussion

The results of Experiment 2 largely replicated and extended the findings from Experiment 1. Once again, in the absence of the JOL queries, the cued-recall interpolated tests produced the largest benefit on new learning in Experiment 2a, with difficult multiple-choice producing a marginal benefit and easy multiple-choice producing a nonsignificant benefit. Moreover, our data showed that interpolated testing lengthened participants' reading time for the criterial section text, and reading time was correlated with performance.

Experiment 2b replicated the results of Experiment 1b, such that when learners were required to repeatedly reflect on their future performance there was no FTE observed. Again, the absence of the FTE in Experiment 2b relative to Experiment 2a appears to be due to an increase in performance in the no-test group ($M_{\mathrm{NT}}=0.47$ in Experiment 2a vs. $M_{\mathrm{NT}\,=}$ 0.56 in Experiment 2b). There was a minimal change for participants in the three tested conditions across the two experiments ($M=0.54$ in Experiment 2a vs. $M=0.56$ in Experiment 2b). Also similar to Experiment 1b, requiring participants to produce JOLs eliminated the differences in reading times across review conditions. Together, these results demonstrate a potentially powerful influence of making JOLs on new learning, which is consistent with metacognitive account of the FTE.

General Discussion

Difficult Tests are Beneficial to New Learning

In the present study, we tested the theoretical tenets of the metacognitive account in an educational context. We manipulated interpolated test difficulty by providing participants with either cued-recall, multiple-choice questions of varying difficulty, or no retrieval opportunities following the reading of text sections. According to the metacognitive account, interpolated tests potentiate new learning because they lead learners to switch to and/or maintain more beneficial learning strategies. Consequently, difficult tests should promote strategy changes more than easy tests, and we should observe a greater FTE when participants take interpolated cued-recall tests than when they take interpolated multiple-choice tests. Indeed, this is what we found in Experiments 1a and 2a. Further, as test difficulty increased from easy multiple-choice to difficult multiple-choice in Experiment 2a, the magnitude of the FTE (as indexed by the effect size of the difference between the tested and non-tested condition) similarly increased.

To further explore how interpolated tests influenced new learning, we combined the data from Experiments 1a and 2a in a random effects meta-analysis using the DerSimonion-Laird method. There was a robust FTE when comparing prior cued-recall to no prior test, $g=0.48$ [0.25, 0.72], $p<0.001$, but only a marginal FTE when comparing prior easy multiple-choice to no prior test, $g=0.21$ [-0.02, 0.44], $p=0.073$. Lastly, performance on the criterial section was also significantly higher following prior cued-recall than prior easy multiple-choice, $g = 0.32$ [0.09, 0.56], $p = 0.007$. This finding is the first of its kind in this literature. By combining the data across Experiments 1a and 2a, we were able detect this more subtle variation in test performance, as our experiments were not powered for this comparison. Although data collection for all of our experiments occurred within three consecutive semesters from the same participant pool at the same University, we add the caveat that readers should exercise caution in interpreting this result given its exploratory nature.

We also showed that interpolated testing increased participants' self-regulated reading times for the criterial text section relative to no interpolated testing, and that reading time was moderately associated with test performance. These results are consistent with the metacognitive account, but we acknowledge that reading time does not provide a direct or exclusive index of strategy use or changes. However, time management and study allocation are essential components of self-regulated learning (Bjork et al., 2013; Kornell & Bjork, 2007; Metcalfe & Kornell, 2005), and do provide a noninvasive indirect measure of how participants' behavior changes in response to interpolated testing[5].

To further investigate this idea, we conducted a mediation analysis in JASP to determine the direct and indirect effects of interpolated tests and reading time using data from Experiments 1 and 3 (where a significant FTE was observed). Prior review type was coded based on the difficulty of the task (no prior test = 1, prior easy multiple choice = 2, prior difficult multiple choice = 3, and prior cued-recall = 4) and served as the predictor variable. Reading time for the criterial section was the mediator, and criterial test performance was the dependent variable, with the bootstrap sample set to 5000. As expected, there was a direct effect of testing on performance, $B = 0.01$ [0.01, 0.02], $p < 0.001$. Most importantly, the indirect path through reading time was also significant, $B = 0.06$ [0.03, 0.11], $p = 0.001$. Thus, reading time is important for improving new learning, but it is not the sole mechanism responsible for the FTE. However, the finding that reading time *does* account for some variation in the indirect path in this model provides converging support for the metacognitive framework. Indeed, the FTE is likely a multi-faceted phenomenon that involves multiple component processes (e.g., Chan et al. 2022; Kliegel & Bäuml, 2021).

Repeated JOLs, Reactivity, and the Elimination of FTE

Unlike Experiments 1a and 2a, where interpolated test difficulty enhanced the forward benefit of testing, a completely different pattern emerged in Experiments 1b and 2b – that is, there was no forward testing effect. Here, participants were asked to make four JOLs after reading each text section (but prior to retrieval). These judgments asked participants to reflect on how much of the just-learned material they would remember for 1) an immediate cued-recall test, 2) an immediate

[5] Further, the most optimal strategy that a learner can use likely varies tremendously based on the difficulty of the to-be-studied material, the learner's motivation, time constraints, and the learner's ultimate goals.

multiple-choice test, 3), a delayed (i.e., in 25 min) cued-recall test, and 4) a delayed multiple-choice test. Despite null performance differences on the criterial test, participants made lower JOLs for cued-recall tests than multiple-choice tests, suggesting that they had carefully considered these questions and had likely engaged in deep metacognitive introspection. This benefit of introspection, in turn, nullified the benefit of retrieval by improving performance in the nontested condition (but not in the tested conditions as the metacognitive knowledge gained through introspection is redundant with that afforded by explicit retrieval practice).

Thus far, only a handful of studies have required learners to make aggregate JOLs in the FTE paradigm (Lee & Ha, 2019; Szpunar et al., 2014; 2013 Yang et al. 2017; Yang et al., 2017). Lee and Ha had participants study two sets of painting-artist pairs, and included three conditions: interpolated testing, interpolated JOLs, and interpolated restudy. They found that asking participants to produce JOLs for the first set of paintings enhanced participants' learning of the second set, similar to the effect of interpolated testing. Unlike the present study, Lee and Ha's participants never completed *both* an interpolated test and JOLs, so it remained possible that producing JOLs and interpolated testing can produce additive benefits on new learning. The current study showed that not to be the case.

Others have also examined how item-by-item JOLs can influence subsequent learning, and whether any reactivity observed can be attributed to covert retrieval, metacognitive reactivity, or both (Soderstrom et al., 2015). Ariel et al. (2021) conducted a thorough investigation of how JOLs influence learning of multiple sections of a text passage. In five experiments, they queried participants for aggregate JOLs (as in the present study) or JOLs for specific concepts. Further, they manipulated the likelihood of covert retrieval during the latter JOL type via the presence or absence of the targeted information during the JOL trial. In all cases, there was no JOL reactivity observed. In contrast, Kubik et al. (2022) found that when participants gave item-level JOLs for cue-target word pairs containing only the cue and the target stem (i.e., when covert retrieval was encouraged), they demonstrated similar benefits on new learning as traditional overt retrieval practice. When the likelihood of covert retrieval was minimized by providing the target in its entirety, new learning did not benefit, suggesting that the JOLs in their study did not induce reactivity over and above the benefits of covert retrieval.

Why did our study and Lee and Ha's (2019) study report JOL reactivity, but others have not? We believe that asking participants to produce JOLs is necessary but not sufficient to induce reactivity. Rather, the JOL task must encourage more than a passing reflection on future performance to have a tangible impact on later studying behavior. Responding to a single question requires little more than assigning a number to a general feeling of knowing that might reflect perceived encoding effort, material difficulty, or general competence (Koriat et al., 2002). Responding to four similar, but slightly different, questions that appear in a random order is likely to require much greater reflections. Indeed, to answer each of our JOL question, participants must take into account the two dimensions of retrieval conditions (e.g., test type and delay) specified in the question and then reflect on how they might affect one's perceived retrieval effort, encoding effort, competence, etc. The finding that participants were sensitive to variations in the judgment questions (e.g., they made higher judgments for multiple-choice questions and for immediate judgments) suggests that participants *did* carefully consider these factors.

Indeed, Lee and Ha (2019) did not find JOL reactivity effects in their Experiment 1, when participants were re-presented with the intact painting-artist pairs during the interpolated phase and asked to provide a JOL for each pair, a task requiring relatively shallow metacognitive processing. It was only in their Experiments 2 and 3, when participants were asked to produce one JOL for each artist's works (here, participants only saw each artist's name so that retrieval might have been required to inform the JOLs), or when participants were given detailed, multi-phase JOL instructions that the metacognitive judgment eliminated the FTE. In a similar vein, Ariel et al. (2021) required participants to make a single JOL after each text section, and this experiment did not include a no-test control. Therefore, we argue that the reactivity effects observed in our experiments may be dependent on the depth or frequency (or both) of processing required by the JOLs.

We also argue that the reactivity effect observed here is not likely due to covert retrieval induced by the metacognitive judgments. There is a wealth of research (for reviews, see Double et al., 2018 and Rhodes, 2016) showing that JOLs can obscure the *backward* effect of retrieval practice. One primary way that JOLs can mask the beneficial backward effect of retrieval practice is that item-by-item JOLs may induce covert retrieval (Spellman & Bjork, 1992; Tauber et al., 2015). This covert retrieval essentially turns a no-testing control into a retrieval condition, improving later test performance, and one might be tempted to argue that the same occurred in the present experiments. However, the likelihood that this type of covert retrieval may have occurred is not high in Experiments 2 and 4, as we required *aggregate* JOLs (i.e., asking to reflect on learning for the entire section, rather than on an item-by-item basis). Lastly, given the amount of information covered in each text section, it is extremely unlikely that participants would be able to guess the four pieces of information on which they would be tested while making the JOLs.

Ariel et al. (2021) found that even under conditions that favor covert retrieval of the exact to-be-tested information, JOLs do not improve performance compared to retrieval alone. While this might seem incongruous with the accounts discussed above, some (Davis & Peterson, 2019; Don et al., 2022) have demonstrated that even partial tests (~50% of items tested) can enhance new learning to the same degree as exhaustive tests. This suggests that it is not item retrieval per se that influences new learning, but rather some more global mechanism that enhances later encoding, retrieval, or some combination of the two.

An alternative, and we believe more probable, possibility is that asking participants to engage in deep and/or repeated metacognitive introspection led them to adopt superior encoding/retrieval strategies as they progressed through the encoding of each section. That is, the metacognitive judgments had downstream consequences similar to testing itself. Rather, by repeatedly thinking back to what they had just read as a whole, participants might produce a global impression of how well they have learned the material and then act accordingly, such as devoting *more time to reading* the later passages, which improved learning particularly for the nontested participants.

It is important to note that participants must be motivated to perform well in order for the information gleaned from providing JOLs to have an effect on behavior. Both intrinsic and extrinsic motivation may be influenced by enhancing learning engagement (ELE; Shi et al, 2023). To illustrate, Shi et al. found that participants who showed reactivity effects reported higher levels of engagement, and that participants showed *reduced* reactivity to JOLs when engagement was increased in another way. While the present study did not explicitly measure motivation or engagement with the material, the interplay between metacognition and endogenous and exogenous motivators may be a potentially fruitful avenue for future research.

In aggregate, these results are consistent with the metacognitive account of the FTE. While future research is needed to determine the extent to which deep metacognitive introspection induces strategy changes, the results from the present series of studies clearly show that requiring JOLs had real and tangible impacts on memory performance in general and the FTE specifically. Taken together, the finding that making JOLs can eliminate the forward benefit of testing support the idea that the formation of expectations of future test performance in general is not always a neutral event but can sometimes induce reflective metacognitive processes that influences future strategy choices.

Unanswered Questions and Limitations

The current set of experiments also introduce some unexpected findings that require further investigation. The most puzzling finding is that the main effect of criterial test type (i.e., superior performance for multiple-choice tests relative to cued-recall tests) observed in Experiments 1a and 2a was eliminated in Experiments 1b and 2b, in which participants were required to make interpolated JOLs. Whereas the elimination of the FTE can be explained by improved performance in the control condition, the lack of a difference between the criterial cued-recall and multiple-choice is more difficult to explain. Put another way, making interpolated JOLs eliminated the difference in difficulty between cued-recall and multiple-choice — a novel finding that we have discovered twice here.

It is ultimately unclear why making aggregate interpolated JOLs would preferentially improve subsequent recall but not multiple-choice performance, but recent research does provide some basis for speculation. Specifically, the occurrence of JOL reactivity may be dependent on the processes involved in making JOLs being reinstated at retrieval. Myers et al. (2020) argued that when studying paired associates, making JOLs would require participants to consider the relationship between the constituents of the pair (i.e., the cue and the target), and when a test requires the reinstatement of this relationship, JOL reactivity (i.e., making JOLs improved retention for the judged pairs) should occur. However, unlike participants in Myers et al., who made item-by-item JOLs during the study phase, participants in our study made retrospective aggregate JOLs; so it remains unclear exactly what type of processes or relationship were involved or reinstated in making the present JOLs. In sum, the puzzling elimination of the difficult effect remains to be investigated in future research.

There were also some differences between the JOL judgments between Experiments 1b and 2b. In Experiment 1b, both prior review type and predicted test type had a significant impact on JOLs. In the former case, participants who had prior cued-recall tests made the lowest JOLs, followed by prior multiple-choice, and participants with no prior test experience made the highest JOLs. Unsurprisingly, the predicted test type main effect showed that participants made higher predictions for future multiple-choice tests relative to future cued-recall tests. In Experiment 2b, this was the only factor that significantly influenced judgments of learning. Aside from the inclusion of the difficult multiple-choice condition (a between-subjects manipulation) and a different sample in Experiment 4, these two experiments were identical. Aside from sampling error, we cannot provide an explanation for this discrepancy.

Some may wonder why we have not suggested that the changes in the JOLs indicate that the reduction in overconfidence is responsible for the FTE.[6] Indeed, we had planned Experiments 1b and 2b as a way to explicitly index metacognitive beliefs underlying strategy changes believed to benefit performance. We are not the first to collect such data, although the few studies that have done so do not always report consistent findings (Lee & Ha, 2019; Szpunar et al., 2014; Yang et al., 2017). Unfortunately, the fact that there was *not* an FTE observed in Experiments 1b and 2b prevents us from using those JOLs to interpret the FTE observed in Experiments 1a and 2a. Precisely how metacognitive beliefs and their calibration to actual performance change as a function of interpolated testing is an important question for future research.

Finally, Yang et al. (2017) and Weinstein et al. (2014) have suggested that the phenomenon of test expectancy could be responsible for the FTE. In this account, participants simply are more likely to *expect* a test to occur when they have been tested recently than when they have not. One might argue that the improvement in performance in the control condition following interpolated JOLs could reflect an increase in test expectancy, rather than a change in strategy per se. We find this explanation problematic for several reasons. First, the test expectancy account does not propose any behavioral changes that might occur as a result of increased test expectancy (Yang et al., 2017). In that sense, the expectancy account would still fall under the metacognitive framework. Moreover, we (Chan et al., 2020; Davis & Peterson, 2019) have found moderate-to-large FTE effects when participants are asked to predict the likelihood of an upcoming test; a procedure which surely would remind participants about the likelihood of impending tests. Finally, there is no a priori reason to assume that participants in the non-tested conditions would interpret repeated metacognitive queries as reminders about upcoming tests, especially when they continue to not be tested.

[6] We would be remiss if we did not add the caveat that the JOL effects that we discuss here were only observed cross-experimentally. In fact, we did not anticipate the JOL reactivity effect at the outset of these experiments, given that there is evidence that reactivity effects are limited in complex materials (Tauber et al., 2015).

Applied Implications and Concluding Remarks

The findings from the current experiments have important applied implications. For educators who administer in-class questions during learning (or students who practice retrieval during textbook reading), the results of Experiments 1a and 2a may suggest that retrieval during learning should generally *not* take the form of multiple-choice questions. Unfortunately, the difficulty associated with scoring recall questions limits their use in some educational settings. However, there were benefits of prior multiple-choice testing relative to no prior testing when the questions were designed to elicit more effortful retrieval. Therefore, our advice for educators who employ tests in the classroom is as follows: teaching without retrieval practice is not ideal, interpolated (difficult) multiple-choice testing is better, and interpolated recall testing is the gold standard whenever possible in educational practice.

However, instructors might be hesitant to insert tests into a class under some circumstances. One might simply have too much material to cover, or fear that students in general dislike tests and could express this distaste with unfavorable teaching evaluations. In this case, our data suggest that asking students to make deep metacognitive reflections might serve as a substitute for testing when considering future learning. Although making JOLs might not replace retrieval practice in terms of its backward benefit, making metacognitive judgments is less time consuming (e.g., instructors do not have to generate the questions, administer the quiz, and score them) and less intimidating than testing. The caveat is that it appears that students must engage in a deep meaningful assessment of their learning for reactivity to benefit future learning, and that shallow judgments are unlikely to confer the same benefit.

To conclude, we have found evidence here that difficult tests lead to enhanced new learning of educational materials (see also Wissman & Rawson, 2015; Wissman et al., 2011). Further, repeated metacognitive queries eliminate the benefit by appearing to improve performance in the control condition(s), a novel finding in the forward testing effect literature. This evidence supports the idea that metacognitive knowledge does contribute to the formation of the FTE, although it is important to remind the reader that the FTE is likely a multi-faceted effect with more than one underlying mechanism, and different mechanisms might be differentially responsible for the effect under different circumstances (Ahn & Chan, 2022, in press; Kliegl & Bäuml, 2021). We believe that future research should continue to investigate the ways in which metacognition can be leveraged to enhance new learning in educational contexts via retrieval practice.

References

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

Abel, M., Bäuml, T., & K. H. (2016). Retrieval practice can eliminate list method directed forgetting. *Memory & Cognition, 44*, 15–23. https://doi.org/10.3758/s13421-015-0539-x

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

Ahn, D., & Chan, J.C.K. (in press). Does testing enhance new learning because it enables learners to use better strategies? *Journal of Experimental Psychology: Learning, Memory, and Cognition*.

Ariel, R., Karpicke, J. D., Witherby, A. E., & Tauber, S. K. (2021). Do judgments of learning directly enhance learning of educational materials? *Educational Psychology Review, 33*, 693–712. https://doi.org/10.1007/s10648-020-09556-8

Bjork, R. A., Dunlosky, J., & Kornell, N. (2013). Self-regulated learning: Beliefs, techniques, and illusions. *Annual Review of Psychology, 64*, 417–444.

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

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

Chan, J. C. K., Meissner, C. A., & Davis, S. D. (2018b). Test-potentiated new learning: A meta-analytic review. *Psychological Bulletin, 114*(11), 1111–1146. https://doi.org/10.1037/bul0000166

Chan, J. C. K., Manley, K. D., & Ahn, D. (2020). Does retrieval potentiate new learning when retrieval stops but new learning continues? *Journal of Memory and Language, 115*, 1–19. https://doi.org/10.1016/j.jml.2020.104150

Chan, J. C., O'Donnell, R., & Manley, K. D. (2022). Warning weakens retrieval-enhanced suggestibility only when it is given shortly after misinformation: The critical importance of timing. *Journal of Experimental Psychology: Applied, 28*(4), 694–716. https://doi.org/10.1037/xap0000394

Cho, K. W., & Powers, A. (2019). Testing enhances both memorization and conceptual learning of categorical materials. *Journal of Applied Research in Memory and Cognition, 8*(2), 166–177. https://doi.org/10.1016/j.jarmac.2019.01.003

Choi, H., & Lee, H. (2020). Knowing Is not half the battle: The role of actual test experience in the forward testing effect. *Educational Psychology Review, 32*(3), 765–789. https://doi.org/10.1007/s10648-020-09518-0

Craik, F. I. (1986). A functional account of age differences in memory. *Human Memory and Cognitive Capabilities: Mechanisms and Performances, 5*, 409–422.

Davis, S. D., Chan, J. C. K., & Wilford, M. M. (2017). The dark side of interpolated testing: Frequent switching between retrieval and encoding impairs new learning. *Journal of Applied Research on Memory and Cognition, 6*, 434–441.

Davis, S. D., & Peterson, D. J. (2019). Reducing the number of retrieval opportunities reduces the magnitude of the forward testing effect. Poster presented at the 60th annual meeting of the Psychonomic Society, Montreal, QC, Canada.

Don, H. J., Yang, C., Boustani, S., & Shanks, D. R. (2022). Do partial and distributed tests enhance new learning? *Journal of Experimental Psychology: Applied, 29*(2), 358–373. https://doi.org/10.1037/xap0000440

Double, K. S., Birney, D. P., & Walker, S. A. (2018). A meta-analysis and systematic review of reactivity to judgments of learning. *Memory, 26*(6), 741–750. https://doi.org/10.1080/09658211.2017.1404111

Dougherty, M. R., Scheck, P., Nelson, T. O., & Narens, L. (2005). Using the past to predict the future. *Memory & Cognition, 33*(6), 1096–1115. https://doi.org/10.3758/BF03193216

Dougherty, M. R., Robey, A. M., & Buttaccio, D. (2018). Do metacognitive judgments alter memory performance beyond the benefits of retrieval practice? A comment on and replication attempt of Dougherty, Scheck, Nelson, and Narens (2005). *Memory & Cognition, 46*, 558–565. https://doi.org/10.3758/s13421-018-0791-y

Dunlosky, J., & Rawson, K. A. (2012). Overconfidence produces underachievement: Inaccurate self evaluations undermine students' learning and retention. *Learning and Instruction, 22*(4), 271–280. https://doi.org/10.1016/j.learninstruc.2011.08.003

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. https://doi.org/10.1177/1529100612453266

Dunn, P. K. (2021). Scientific research and methodology: An introduction to quantitative research in science and health. https://bookdown.org/pkaldunn/Book

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

Finn, B., & Roediger, H. L., III. (2013). Interfering effects of retrieval in learning new information. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 39*(6), 1665–1681. https://doi.org/10.1037/a0032377

Finn, B. (2017). A framework of episodic updating: An account of memory updating after retrieval. In Psychology of learning and motivation. *Academic Press, 67*, 173–211.

Gierl, M. J., Bulut, O., Guo, Q., & Zhang, X. (2017). Developing, analyzing, and using distractors for multiple-choice tests in education: A comprehensive review. *Review of Educational Research, 87*(6), 1082–1116. https://doi.org/10.3102/0034654317726529

Hayslip, B., Jr., & Kennelly, K. J. (1982). Short-term memory and crystallized-fluid intelligence in adulthood. *Research on Aging, 4*(3), 314–332. https://doi.org/10.1177/0164027582004003003

Janes, J. L., Rivers, M. L., & Dunlosky, J. (2018). The influence of making judgments of learning on memory performance: Positive, negative, or both? *Psychonomic Bulletin & Review, 25*(6), 2356–2364.

JASP Team, (2023). JASP (Version 0.17). [Computer software].

Jing, H. G., Szpunar, K. K., & Schacter, D. L. (2016). Interpolated testing influences focused attention and improves integration of information during a video-recorded lecture. *Journal of Experimental Psychology: Applied, 22*, 305–318. https://doi.org/10.1037/xap0000087

King, J. F., Zechmeister, E. B., & Shaughnessy, J. J. (1980). Judgments of knowing: The influence of retrieval practice. *The American Journal of Psychology,* 329–343.

Kliegl, O., & Bäuml, K.-H.T. (2021). When retrieval practice promotes new learning – The critical role of study material. *Journal of Memory and Language, 120*, 104253. https://doi.org/10.1016/j.jml.2021.104253

Koriat, A., Sheffer, L., & Ma'ayan, H. (2002). Comparing objective and subjective learning curves: Judgments of learning exhibit increased underconfidence with practice. *Journal of Experimental Psychology: General, 131*(2), 147–162.

Kornell, N., & Metcalfe, J. (2006). Study efficacy and the region of proximal learning framework. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 32*(3), 609–622. https://doi.org/10.1037/0278-7393.32.3.609

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

Kubik, V., Koslowski, K., Schubert, T., & Aslan, A. (2022). Metacognitive judgments can potentiate new learning: The role of covert retrieval. *Metacognition and Learning, 17*, 1057–1077. https://doi.org/10.1007/s11409-022-09307-w

Lee, H. S., & Ha, H. (2019). Metacognitive judgments of prior material facilitate the learning of new material: The forward effect of metacognitive judgments in inductive learning. *Journal of Educational Psychology, 111*(7), 1189–1201. https://doi.org/10.1037/edu0000339

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. https://doi.org/10.1177/0956797612443370

Mazzoni, G., Cornoldi, C., & Marchitelli, G. (1990). Do memorability ratings affect study-time allocation? *Memory & Cognition, 18*(2), 196–204. https://doi.org/10.3758/BF03197095

Metcalfe, J., & Kornell, N. (2005). A region of proximal learning model of study time allocation. *Journal of memory and language, 52*(4), 463–477. https://doi.org/10.1016/j.jml.2004.12.001

McDaniel, M. A., Anderson, J. L., Derbish, M. H., & Morrisette, N. (2007). Testing the testing effect in the classroom. *European Journal of Cognitive Psychology, 19*(4–5), 494–513. https://doi.org/10.1080/09541440701326154

McDermott, K. B. (2021). Practicing retrieval facilitates learning. *Annual Review of Psychology, 72*, 609–633. https://doi.org/10.1146/annurev-psych-010419-051019

Mitchum, A. L., Kelley, C. M., & Fox, M. C. (2016). When asking the question changes the ultimate answer: Metamemory judgments change memory. *Journal of Experimental Psychology: General, 145*(2), 200–219. https://doi.org/10.1037/a0039923

Morris, C. D., Bransford, J. D., & Franks, J. J. (1977). Levels of processing versus transfer appropriate processing. *Journal of Verbal Learning and Verbal Behavior, 16*(5), 519–533. https://doi.org/10.1016/S0022-5371(77)80016-9

Myers, S. J., Rhodes, M. G., & Hausman, H. E. (2020). Judgments of learning (JOLs) selectively improve memory depending on the type of test. *Memory and Cognition, 48*, 745–758. https://doi.org/10.3758/s13421-020-01025-5

Panadero, E., Jonsson, A., & Botella, J. (2017). Effects of self-assessment on self-regulated learning and self-efficacy: Four meta-analyses. *Educational Research Review, 22*, 74–98. https://doi.org/10.1016/j.edurev.2017.08.004

Pastötter, B., & Bäuml, K. T. (2014). Retrieval practice enhances new learning: The forward effect of testing. *Frontiers in Psychology, 5*, 5.

Pastötter, B., Schicker, S., Niedernhuber, J., & Bäuml, K. H. T. (2011). Retrieval during learning facilitates subsequent memory encoding. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 37*, 287–297. https://doi.org/10.1037/a0021801

Pastötter, B., Weber, J., & Bäuml, K. H. T. (2013). Using testing to improve learning after severe traumatic brain injury. *Neuropsychology, 27*(2), 280–285. https://doi.org/10.1037/a0031797

Pastötter, B., Engel, M., & Frings, C. (2018). The forward effect of testing: Behavioral evidence for the reset-of-encoding hypothesis using serial position analysis. *Frontiers in Psychology, 9*, 1197. https://doi.org/10.3389/fpsyg.2018.01197

Rawson, K. A. (2015). The status of the testing effect for complex materials: Still a winner. *Educational Psychology Review, 27*(2), 327–331. https://doi.org/10.1007/s10648-015-9308-4

Rhodes, M. G. (2016). Judgments of learning: Methods, data, and theory. In J. Dunlosky, & S. K. Tauber (Eds.), The Oxford Handbook of Metamemory; (pp. 65–80, Chapter xv, 574 Pages) Oxford University Press, New York, NY.

Roediger, H. L., III., & Marsh, E. J. (2005). The positive and negative consequences of multiple-choice testing. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 31*(5), 1155–1159. https://doi.org/10.1037/0278-7393.31.5.1155

Rowland, C. A. (2014). The effect of testing versus restudy on retention: A meta-analytic review of the testing effect. *Psychological Bulletin, 140*(6), 1432–1463. https://doi.org/10.1037/a0037559

Sahakyan, L., Delaney, P. F., & Kelley, C. M. (2004). Self-evaluation as a moderating factor of strategy change in directed forgetting benefits. *Psychonomic Bulletin & Review, 11*(1), 131–136. https://doi.org/10.3758/BF03206472

Shi, A., Xu, C., Zhao, W., Shanks, D. R., Hu, X., Luo, L., & Yang, C. (2023). Judgments of learning reactively facilitate visual memory by enhancing learning engagement. *Psychonomic Bulletin & Review, 30*(2), 676–687. https://doi.org/10.3758/s13423-022-02174-1

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

Soderstrom, N. C., Clark, C. T., Halamish, V., & Bjork, E. L. (2015). Judgments of learning as memory modifiers. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 41*(2), 553–558. https://doi.org/10.1037/a0038388

Spellman, B. A., & Bjork, R. A. (1992). When predictions create reality: Judgments of learning may alter what they are intended to assess. *Psychological Science, 3*(5), 315–317. https://doi.org/10.1111/j.1467-9280.1992.tb00680.x

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

Szpunar, K. K., Khan, N. Y., & Schacter, D. L. (2013). Interpolated memory tests reduce mind wandering and improve learning of online lectures. *Proceedings of the National Academy of Sciences of the United States of America, 110*, 6313–6317.

Szpunar, K. K., Jing, H. G., & Schacter, D. L. (2014). Overcoming overconfidence in learning from video-recorded lectures: Implications of interpolated testing for online education. *Journal of Applied Research in Memory and Cognition, 3*, 161–164.

Tauber, S. K. U., Dunlosky, J., & Rawson, K. A. (2015). The influence of retrieval practice versus delayed judgments of learning on memory: Resolving a memory-metamemory paradox. *Experimental Psychology, 62*(4), 254–263. https://doi.org/10.1027/1618-3169/a000296

Thiede, K. W., & Dunlosky, J. (1994). Delaying students' metacognitive monitoring improves their accuracy in predicting their recognition performance. *Journal of Educational Psychology, 86*(2), 290–302. https://doi.org/10.1037/0022-0663.86.2.290

Weinstein, Y., Gilmore, A. W., Szpunar, K. K., & McDermott, K. B. (2014). The role of test expectancy in the build-up of proactive interference in long-term memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 40*(4), 1039–1048. https://doi.org/10.1037/a0036164

Wissman, K. T., & Rawson, K. A. (2015). Grain size of recall practice for lengthy text material: Fragile and mysterious effects on memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 41*, 439–455. https://doi.org/10.1037/xlm0000047

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*, 1140–1147.

Woodford, C. (2021). Lasers. Explain that Stuff. Retrieved January 24, 2023 from https://www.explainthatstuff.com/lasers.html

Yang, C., Potts, R., & Shanks, D. R. (2017). The forward testing effect on self-regulated study time allocation and metamemory monitoring. *Journal of Experimental Psychology: Applied, 23*(3), 263–277. https://doi.org/10.1037/xap0000122

Yang, C., Potts, R., & Shanks, D. R. (2018). Enhancing learning and retrieval of new information: a review of the forward testing effect. *NPJ science of learning, 3*(1), 8. https://doi.org/10.1038/s41539-018-0024-y

Yang, C., Chew, S. J., Sun, B., & Shanks, D. R. (2019). The forward effects of testing transfer to different domains of learning. *Journal of Educational Psychology, 111*(5), 809–826. https://doi.org/10.1037/edu0000320

Yang, C., Zhao, W., Luo, L., Sun, B., Potts, R., & Shanks, D. R. (2022). Testing potential mechanisms underlying test-potentiated new learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 48*(8), 1127–1143. https://doi.org/10.1037/xlm0001021

Zhao, W., Li, B., Shanks, D. R., Zhao, W., Zheng, J., Hu, X., Su, N., Fan, T., Yin, Y., Luo, L., & Yang, C. (2021). When judging what you know changes what you really know: Soliciting metamemory judgments reactively enhances children's learning. *Child Development, 93*(2), 405–417. https://doi.org/10.1111/cdev.13689

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