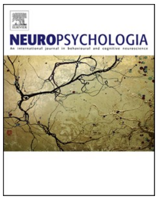
Neuropsychologia 138 (2020) 107298

Contents lists available at ScienceDirect

Neuropsychologia

journal homepage: http://www.elsevier.com/locate/neuropsychologia



A role for familiarity in supporting the testing effect over time

Ruth A. Shaffer[a,*], Kathleen B. McDermott[a,b]

[a] Department of Psychological & Brain Sciences, Washington University in St. Louis, St. Louis, MO, USA
[b] Department of Radiology, Washington University School of Medicine, St. Louis, MO, USA

ARTICLE INFO

Keywords:
Dual process
Retrieval practice
Recognition
Delay
Recollection
Familiarity

ABSTRACT

Endel Tulving (1985) drew a distinction between Remembering and Knowing, spurring a great deal of research on the memorial experiences of recollection and familiarity and their contribution to various phenomena in memory. More recently, studies have used this distinction to situate our understanding of the processes that contribute to the testing effect—or, the benefit of retrieval practice to later memory (see also Tulving, 1967). Using retention intervals of approximately 15 min or less between initial and final testing, several studies have found that initial testing magnifies estimates of recollection but not familiarity, regardless of whether a testing effect is revealed in overall recognition performance (Chan and McDermott, 2007). However, the efficacy of prior testing in enhancing memory has been shown to change over time, as have estimates of recollection and familiarity. Thus, the mechanisms that underlie the quintessential testing effect—one that occurs in overall recognition or recall over longer delays—are still uncertain. To investigate this issue, in two experiments, subjects studied word lists, took 3-letter stem cued-recall tests on half of the studied words, and completed a final recognition test in which estimates of recollection and familiarity were obtained via confidence (Experiment 1) or Remember-Know-New (Experiment 2) judgments. Critically, final recognition tests occurred either immediately, 1 day (Experiment 1 only), or 4 days after initial learning. At all retention intervals and in both methods of estimating recollection and familiarity on the final test (i.e. receiver-operating characteristic and remember-know analyses), initial testing magnified estimates of both recollection and familiarity. These findings suggest that the testing effect can result from changes in both processes and pose issues for theories of the testing effect that consider an exclusive role for recollection.

1. Introduction

1.1. The testing effect

> "In very short order we lose something like 70 percent of what we've just heard or read. After that, forgetting begins to slow … but the lesson is clear: a central challenge to improving the way we learn is finding a way to interrupt the process of forgetting" (Brown et al., 2014, p. 28).

Which learning strategies lead to the greatest long-term retention? Researchers have asked this question for years. And at the forefront of recent discussion has been the testing effect. The testing effect refers to the finding that taking a test does not simply serve as a passive indication of what a person knows. Indeed, taking a test, or retrieval practice, has been shown to enhance long-term retention of material relative to not retrieving and even to rereading material (Dunlosky et al., 2013; Karpicke and Roediger, 2008; Roediger and Butler, 2011; Roediger and Karpicke, 2006a; Rowland, 2014; Tulving, 1967). Although a majority of the testing effect literature engages participants in a laboratory setting, a significant, albeit smaller, literature has also provided strong evidence for testing's benefit in the classroom (e.g., Larsen et al., 2008; McDaniel et al., 2013; McDermott et al., 2014; Roediger et al., 2011). Thus, the fact *that* retrieval practice can lead to marked gains in long term retention is well established; however, *how and why* this is the case is much less understood. The experiments reported in this paper seek to better understand the processes that lead to the benefits of testing to later memory.

1.2. Recollection and familiarity in the dual-process perspective

Important to this discussion is an understanding of recollection and

[*] Corresponding author. Department of Psychological & Brain Sciences, Campus Box 1125, One Brookings Drive, Washington University in St. Louis, St. Louis, MO, 63130, USA.
E-mail address: ruthieshaffer@wustl.edu (R.A. Shaffer).

https://doi.org/10.1016/j.neuropsychologia.2019.107298
Received 30 June 2019; Received in revised form 30 November 2019; Accepted 5 December 2019
Available online 12 December 2019
0028-3932/© 2019 Published by Elsevier Ltd.

familiarity, two processes thought to be involved in making memorial decisions during tests (Chan and McDermott, 2007; Jacoby, 1991; Tulving, 1985; Yonelinas, 2002). Endel Tulving (1985) drew an initial distinction between "Remembering" and "Knowing," prompting a wave of research on the memorial experiences of what would later be referred to as recollection and familiarity. In his framework, Tulving proposed the existence of a system of memory accompanied by conscious recollection of the experiences from which the memory originated (evidenced via "Remember" responses in the Remember-Know procedure). He proposed the existence of a separate system of memory for information unaccompanied by this conscious recollection (evidenced via "Know" responses and originally conceptualized as an index of what Tulving referred to as semantic memory).

Ultimately, the distinction between recollection and familiarity was further defined in the literature (Yonelinas, 2001, 2002). Indeed, from the dual-process perspective, recollection is characterized by "effortful" remembering and feeling as though one "re-experiences" an earlier encounter with the tested material. By contrast, familiarity is characterized by decontextualized memory and an inclination toward (or, to risk a tautology, "familiarity" with) material without "re-experiencing" the initial encounter (Jacoby, 1991; Yonelinas, 2002). A rich literature has since distinguished the two processes from one another with respect to various dimensions, such as their differential use on recall- and recognition-based tests, their persistence over short and long delays, their change across age, and their neural correlates, among other factors (Gardiner and Java, 1991; Koen and Yonelinas, 2016; Skinner and Fernandes, 2007; Yonelinas and Levy, 2002; for review, see Yonelinas, 2002). However, here we contend that the way in which recollection and familiarity processes support the benefit of retrieval practice to later memory—or, the way in which prior testing affects these two processes—is still unclear.

1.3. Applying the dual-process perspective to the testing effect

This is not to say that the topic has been completely ignored. Chan and McDermott (2007) conducted three experiments designed to explore whether initial retrieval practice might facilitate subsequent successful recollection, even when overall recognition performance (i.e., the hit rate) does not reveal a testing effect. In their study, participants studied a list of words and either took recall tests on previously studied material or solved math problems. Estimates of recollection and familiarity were calculated on a final recognition test taken up to 15 min later as a function of prior learning condition (study with an initial recall test or study with a distractor math task). Using three canonical methods for estimating recollection and familiarity (source memory, Remember-Know, and exclusion tests), the authors found that prior testing did magnify estimates of recollection on the final test but did not affect the use of familiarity (also see Jones and Roediger, 1995). Other work using a restudy, rather than a no test, control condition corroborated this finding (Verkoeijen et al., 2011; Pu and Tse, 2014; Rowland, 2011; although see Gao et al., 2016).

On the surface, these results lead to the conclusion that the mechanisms responsible for the benefits of retrieval practice to later memory must lie in recollective—and not familiarity—processes. Indeed, theories of the testing effect often center around recollection-related processes. For example, one framework holds that retrieval practice improves memory via enhanced recollection for temporal context information (Karpicke et al., 2014; Lehman et al., 2014). Another conceptualization argues that elaboration processes that occur during initial testing are responsible for the benefits of testing (Carpenter, 2009; Carpenter and DeLosh, 2006). However, it may be premature to conclude that the benefit of retrieval practice accrues exclusively through enhanced recollection. We believe that a critical element is missing from much of the research reported thus far. And that element is delay.

1.4. The impact of delay

Although the testing effect emerges in various situations, critical to its reliable appearance is delay, especially when the initial test occurs without feedback. When the final test occurs immediately following initial testing, the characteristic testing effect may disappear, or even reverse. Completing retrieval practice, relative to completing an unrelated distractor task, may not enhance final test performance after only a short delay. Further, *rereading*, relative to completing retrieval practice, may lead to better final test performance. After one to two days delay, however, prior testing reliably improves performance (Roediger and Butler, 2011; Roediger and Karpicke, 2006a, 2006b). Critically, thus far much of the literature examining the mechanisms that underlie the effect from the dual process perspective have used short delays of about 15 min or less between initial and final tests (although see Bies-Hernandez, 2013). This situation leaves unresolved the mechanisms that are responsible for the quintessential testing effect—one that occurs over retention intervals of significantly greater lengths than 15 min.

1.5. A potential role for familiarity

Several findings point to the possibility that familiarity processes may contribute differentially to the testing effect over time. Indeed, just as the memorial benefits of retrieval practice change over time, the relative contributions of recollection and familiarity to memory are non-static. Immediately after initial exposure, familiarity declines more rapidly than does recollection. Across longer delays, however, familiarity declines at a similar or even slower rate than does recollection (Gardiner and Java, 1991; Yonelinas, 2002). The so-called "Remember-to-Know" shift also casts doubt on the idea that familiarity must play no role in the delayed testing effect. Specifically, the "Remember-to-Know" shift refers to the finding that, on successive tests, many items that initially are classified as "remembered" (i.e., recollected) are given "know" or "familiar" responses on later tests (Dewhurst et al., 2009; also see Conway et al., 1997). Thus, over time, what was initially enhanced recollection may become decontextualized and appear as enhanced familiarity on a delayed final test.

The effect of testing on delayed source memory judgments casts further doubt on the assertion that recollection alone supports the benefit of testing. Whereas recognition with correct source memory is thought to be indicative of recollection, recognition with incorrect source memory is thought to be indicative of familiarity in the absence of recollection. Calculations made by the present authors from group means reported in Dudukovic et al. (2009) Exp. 1 and Kessler et al. (2014) indicate that prior testing numerically elevates the overall proportion of correct *and incorrect* delayed source memory responses.[1] This finding stands in contrast to Chan and McDermott (2007), who found, after a 15 min retention interval, that prior testing improved correct source memory but actually *reduced* the proportion of incorrect source memory responses. Notably, however, the experiments conducted by Dudukovic et al. (2009) and Kessler et al. (2014) were not designed with the aim of addressing this question. As such, these results should be treated as merely suggestive that familiarity processes can contribute to the delayed testing effect.

The most direct evidence for familiarity's involvement in supporting the testing effect over longer retention intervals derives from Bies-Hernandez (2013), who in her dissertation examined the effect of initial testing (relative to restudying) on retention of material after a 2-day delay. Results indicated that prior testing magnified delayed

[1] Calculations made by the present authors from group means reported in Kessler et al. (2014) additionally reveal this pattern of results on a final test that immediately follows the initial test. However, initial testing occurred 1 day after initial studying. Thus, it is unclear the extent to which the result is directly comparable to that of Chan and McDermott (2007)'s immediate final test.

estimates of recollection after multiple—but not one—initial tests. Critically, prior testing magnified estimates of familiarity in all cases, contradicting the majority of prior work using short retention intervals. Several design choices, however, made comparison with prior work difficult (e.g., the use of a 2-day retention interval only, a fully between-subjects design, and a restudy with feedback control condition). This situation leaves open to question the influence of retention interval, specifically, on the apparent contribution of familiarity to the testing effect.

Beyond the impact of retention interval, several findings suggest that, even in the short-term, familiarity may play a role in producing the benefits of retrieval practice. For example, contrary to the majority of work using short retention intervals, one study found that the magnitude of an observed positive testing effect did not differ for estimates of recollection and familiarity (Gao et al., 2016). Further evidence for a role for familiarity in the immediate testing effect has come from studies of aging and of divided attention. For example, although recollection has been found to be impaired and familiarity preserved in older relative to younger adults (Anderson et al., 2008; Bastin and Van der Linden, 2003; Koen and Yonelinas, 2016; Yonelinas, 2002), older adults have shown benefits of retrieval practice equal to that of younger adults on both immediate and delayed final tests (Coane, 2013; Logan and Balota, 2008; Rabinowitz and Craik, 1986, but see Tse et al., 2010). In addition, divided attention during retrieval has been shown to impair recollective processing and leave familiarity relatively preserved (Dudukovic et al., 2009; Jacoby, 1991; Yonelinas, 2002). However, several studies find similar (or even improved) final test performance following conditions in which attention is divided during initial testing relative to full attention conditions on both immediate (Kessler et al., 2014; Mulligan and Picklesimer, 2016) and delayed final tests (Gaspelin et al., 2013; Kessler et al., 2014; Mulligan and Picklesimer, 2016; but see Buchin and Mulligan, 2017; Dudukovic et al., 2009). Thus, even when initial recollective processing is impaired, a benefit of prior testing can be observed similar to that observed under conditions in which initial recollective processing is not impaired. This outcome suggests that, in some cases, familiarity processes may help drive even the immediate testing effect.

Although the above findings leave open the possibility that familiarity can contribute to the benefits of retrieval after short delays, stronger is the suggestion that the *long-term* benefits of retrieval practice may be influenced by changes in familiarity. However, theoretical discussions and studies of the testing effect often refer to the effect as a recollection-only phenomenon. Many studies report changes in recollective processes (e.g. source memory) following retrieval practice, but they do not provide a means of assessing changes in familiarity processes that may likewise support the effect. Thus, it is important to determine the validity and reliability of the assertion that recollection processes alone support the short-term and long-term benefits of retrieval practice.

1.6. Neural mechanisms of the testing effect

To date, a small number of studies have examined the neural mechanisms that underlie the testing effect, and the results have been largely inconclusive (see van den Broek et al., 2016, for a review). Functional magnetic resonance imaging (fMRI) studies on the effects of repeated testing have pointed to a variety of regions as potential correlates of the testing effect, such as the anterior cingulate cortex (ACC) (Eriksson et al., 2011), inferior frontal gyrus (IFG), ventrolateral and dorsolateral prefrontal cortex (VLPFC; DLPFC), precuneus, and inferior parietal lobule (IPL) (Hashimoto et al., 2011). Studies comparing the effects of testing to restudying have implicated some of the same regions in the effect, along with a wide variety of other regions and networks throughout the brain, such as the Default Mode Network and a network of regions implicated in working memory (Keresztes et al., 2014; Liu et al., 2014; van den Broek et al., 2013; Wing et al., 2013; see van den Broek et al., 2016, for a recent review).

Interpretation of these neuroimaging findings has been similarly varied. Some have considered context reinstatement and search set restriction processes that relate to the episodic context account (Gao et al., 2016; Keresztes et al., 2014; Liu et al., 2017; Peng et al., 2019; van den Broek et al., 2013), a prominent theory of the testing effect. Others have considered elaboration processes related to the semantic elaboration account (Liu et al., 2014; Rosburg et al., 2015; Wing et al., 2013), another prominent theory of the effect. However, interpretation has also included a role in the testing effect for consolidation processes (Eriksson et al., 2011; Wing et al., 2013), cognitive control and recollection-related processes broadly (Hashimoto et al., 2011; van den Broek et al., 2013), retrieval monitoring processes (Hashimoto et al., 2011), and Transfer-Appropriate Processing (Rosburg et al., 2015). In a review of the extant neuroimaging literature on the topic, van den Broek et al. (2016) argued that the imaging findings suggest that multiple mechanisms (e.g. search set restriction and elaboration processes) may together contribute to the effect.

To our knowledge, all fMRI investigations of the testing effect have used cued-recall, rather than recognition, tests to assess final memory performance. Notably, performance on cued-recall tests is thought to rely more heavily on the use of recollection than performance on recognition tests (Yonelinas, 2002). In the current study we employ final recognition testing, which encourages the possibility that a benefit of testing to familiarity processes, if truly present, will be observed on the final test. Research on the neural correlates of recollection and familiarity during *recognition* testing, specifically, has suggested that recollection may be preferentially supported by regions in the inferior lateral parietal cortex, the hippocampus, and the parahippocampal cortex (PhC) (Diana et al., 2013; Ranganath, 2010; Vilberg and Rugg, 2007; Yonelinas et al., 2005). By contrast, regions in the superior lateral parietal cortex and perirhinal cortex (PrC) have been associated in part with item familiarity (Ranganath, 2010; Vilberg and Rugg, 2007; but see Diana et al., 2013). In the present study, behavioral indices of recollection and familiarity obtained on immediate and delayed recognition tests can aid in providing predictions for future research as to the neural mechanisms that may support the testing effect over time.

1.7. The present study

The current study was designed with the goal of examining the extent to which changes in recollection and familiarity processes support the benefits of retrieval practice over time—from an immediate to a four-day delayed final test. In two experiments, participants studied words, completed 3-letter stem cued recall tests on half of the words, and took a final recognition test either immediately, 1 day later (Exp. 1 only), or 4 days later. On the final recognition test, participants discriminated between old and new words and indicated their confidence in their response (Exp. 1) or provided a Remember-Know judgment (Exp. 2). We hypothesize that with increasing delay, the testing effect would be supported by changes in both recollection and familiarity processes.

2. Experiment 1

The first aim of Experiment 1 was to conceptually replicate prior work establishing the processes that underlie the benefit of prior testing on an immediate final test. The second aim was to examine the extent to which findings regarding the processes that support the testing effect on an immediate final test are altered or maintained after a 1-day and 4-day retention interval.

2.1. Materials and methods

2.1.1. Participants

One-hundred and fifteen subjects completed Session 1 of the experiment on Amazon Mechanical Turk (immediate final test condition: 37; 1-day delay: 38; 4-day delay: 40). Of these, 2 subjects were excluded because they failed to complete Session 2 (both from the 4-day delay condition). An additional 22 subjects were excluded due to a combination of the following reasons: noting down words during the task (10), restarting the experiment after beginning the main task (2), missing data (4; e.g. missing final test trials), reporting non-normal or corrected-to-normal vision (5), reporting a possible neurological disorder (4), and reporting that English is not the subject's native language (1). One additional subject was excluded due to extreme familiarity estimates (more than 5 standard deviations below the mean). In addition, subjects were excluded if responses were made in under 250 ms to 5 or more items during the final test. An additional 7 subjects were excluded for this reason (immediate: 1; 1-day delay: 4; 4-day delay: 2). All subjects were at least 18 years of age. Eighty-three subjects met all inclusion criteria: immediate final test (N = 29, mean age (years) = 36.9, SD age = 11.0, age range = 23–61, female [F] = 15), 1-day delay (N = 25, mean age = 37.7, SD age = 9.6, age range = 27–59, F = 9), 4-day delay (N = 29, mean age = 34.8, SD age = 8.7, age range = 23–53, F = 16). Conditions were run sequentially on Amazon Mechanical Turk (1: immediate; 2: 1-day delay; 3: 4-day delay). A minimum sample size of 15 participants per condition was determined via an a priori power analysis of Chan and McDermott (2007) Exp. 3 (reported effect size of *t*-test examining recollection estimates = 0.80, alpha = 0.05, and desired power = 0.8), which was achieved in each case (power analysis was conducted using G*Power).

Subjects were compensated approximately $9.00 per hour for their participation. Before beginning the study, subjects consented to participation and, upon completion of the study, were debriefed. The experiment was conducted in accordance with the Washington University in St. Louis Institutional Review Board.

2.1.2. Materials

Stimuli consisted of 240 words and were obtained from the English Lexicon Project database (Balota et al., 2007). Stimulus selection parameters were chosen to emulate those of Chan and McDermott (2007; Experiment 3): Kučera-Francis Frequency = 5–200; parts of speech = noun, adjective, and/or verb; word length = 4–9 letters. Words were selected so that none had the same first 3 letters and so that average frequency constraints would be met based on Chan and McDermott (2007). Specifically, 8 lists of 30 words were constructed so that each list's average Kučera-Francis Frequency fell within 34.33–35. List placement within the experiment was counterbalanced in an effort to place each list in the Test, No Test, and New (on the final test) conditions approximately 25%, 25%, and 50% of the time, respectively. The order of a given stimulus within an experimental block to which it was assigned was newly randomized for each subject and session.

2.1.3. Procedure

The experiment included an Initial Learning session, in which subjects studied words and took 3-letter stem cued-recall tests, and a Final Testing session, in which memory for previously studied, tested, and new items was probed on a recognition test. The procedure is based on Chan and McDermott (2007), Experiment 3, and is detailed below for each session (see Fig. 1 for a depiction of the experimental design).

2.1.3.1. Initial learning. Initial Learning condition (Test vs. No Test) was manipulated within subjects and was divided into two blocks of study-test cycles in order to improve initial test performance. Subjects were instructed that during the experiment they would be shown a set of words and asked to remember the words for a later memory test.

In the first Initial Learning block, 60 words were visually presented sequentially for 5s each (1s ISI). Subjects then took a cued-recall test on half of the words from the preceding list (30 words), in which they were shown the first 3 letters of a word and were given 7s to type the complete word in the blank provided (or to leave it empty if they could not remember the previously studied word; 1s ISI). In the second block, subjects completed an identical study session with a new set of 60 words and an identical cued-recall test on half of the new set of 60 words. No feedback was given during the cued-recall tests.

2.1.3.2. Retention interval. Subjects completed the Final Testing session either 1) immediately (average retention interval between Initial Learning and Final Testing = 18.00 s, SD = 14.40 s; range = 8.40 s–69.00 s); 2) approximately 1 day later (average retention interval = 1.10 days, SD = 0.13 days, range = 0.96 days–1.43 days); or 3) approximately 4 days later (average retention interval = 4.50 days, SD = 0.73 days, range = 3.95 days–6.38 days).

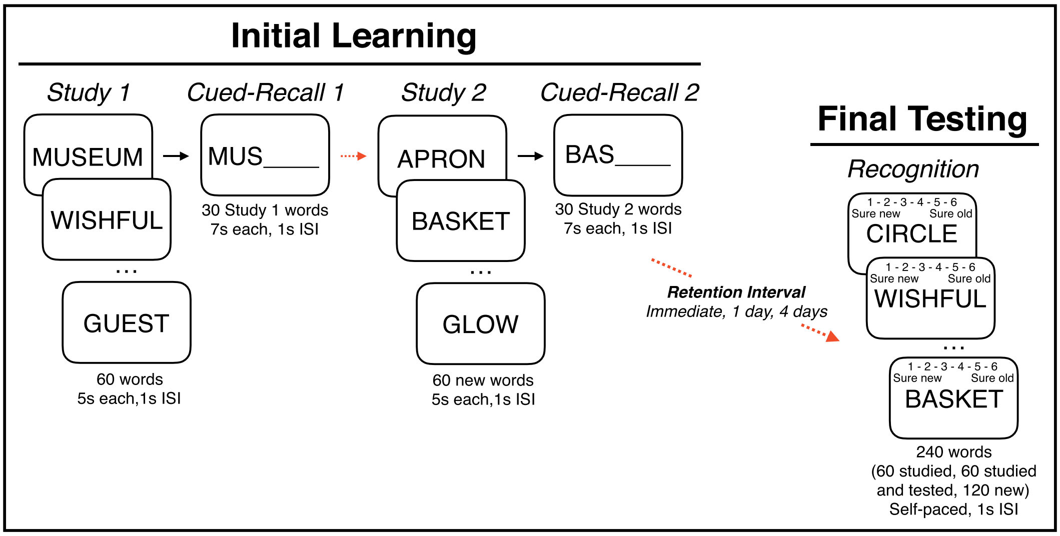


Fig. 1. Design of Experiment 1. Experiment 1 consisted of an Initial Learning stage and a Final Testing stage. Initial Learning was divided into two blocks in order to enhance initial test accuracy. During Initial Learning, subjects 1) studied 60 words for 5s each (1s ISI); 2) took a 3-letter stem cued-recall test on half of the studied words (7s per response, 1s ISI); 3) studied another set of 60 words for 5s each (1s ISI); and 4) took another 3-letter stem cued-recall test on half of the newly studied words (7s per response, 1s ISI). Subjects then took a final test either immediately, 1 day later, or 4 days later. Specifically, during the final test subjects completed a self-paced recognition test that included all of the old words and an equal number of new words and indicated whether they believed each stimulus to be old or new on a scale from 1 to 6 (1 = sure new and 6 = sure old, with varying levels of confidence in between).

2.1.3.3. Final testing. During the final recognition test, subjects were shown all 120 words they had previously studied (half of which had also been tested), and an equal number of new words (120 words). For each word presented, subjects were asked to indicate whether it was old or new to the experiment using a confidence scale from 1 to 6 (1 = sure new, 2 = maybe new, 3 = guess new, 4 = guess old, 5 = maybe old, 6 = sure old). These confidence data were then fit to the Dual-Process Signal Detection model via an ROC Toolbox (Koen et al., 2017; MATLAB ROC Toolbox Version 1.1.3) to obtain estimates of recollection and familiarity on the final test (Yonelinas, 1994; Yonelinas et al., 2010; Yonelinas and Parks, 2007).

Lastly, subjects answered several questions about their experience during the task (e.g. whether they noted down any of the words) and provided demographic information.

2.2. Results

For a given statistical analysis, if any outliers were detected (more than 3 standard deviations above or below the mean), a duplicate analysis excluding outliers was conducted. When outliers were detected, no differences were observed in the patterns of results.

2.2.1. Initial learning

Initial test performance was first analyzed to ensure that there were no significant pre-existing differences in initial learning across the three retention interval groups. Accuracy during Initial Learning was calculated for each subject as proportion correct on the two initial cued-recall tests. For 2 subjects, performance on 1 of the 60 initial cued-recall test trials failed to log. For these 2 subjects, initial accuracy was calculated as proportion correct out of 1 fewer trials. Items were scored leniently. For example, if the correct answer was “OFFERED,” responses such as “OFFER” and “OFFERING” (along with misspellings, such as “OFFERRED”) were counted as correct.

As can be seen in Table 1, participants in the three retention interval groups performed similarly on the initial test, as expected given that the delay manipulation occurred after the initial test. In addition, performance improved from the first to the second initial cued-recall test, likely reflecting practice effects. This improvement, however, did not significantly differ for the three retention interval groups.

A Two-Way Mixed ANOVA formally examined the effect of retention interval group (between-subjects: immediate, 1 day delay, 4 day delay) and test block (within-subjects: initial test block 1, initial test block 2) on initial test accuracy. The main effect of retention interval was not significant ($F(2, 80) = 0.38, p = .684, \eta_p^2 = .01$); subjects in the three delay conditions were similar in overall initial test performance ($M_{\text{immediate}} = .51$, $M_{1\text{ day}} = .49$, $M_{4\text{ day}} = .47$). There was, however, a significant main effect of initial test block; cued-recall accuracy increased from the first ($M = .44$) to the second ($M = .53$) test ($F(1, 80) = 23.31, p < .001, \eta_p^2 = .23$). The interaction between retention interval group and initial test block, however, was not significant ($F(2, 80) = 1.08, p = .346, \eta_p^2 = .03$).

Table 1
Initial cued-recall test performance in Experiment 1. Means (and standard errors) are reported for initial test blocks 1 and 2 by retention interval group (immediate, 1 day delayed, 4 day delayed).

Experiment 1: Initial Cued-Recall Test Accuracy		
Retention Interval	Initial Test Block 1	Initial Test Block 2
Immediate	.44 (.03)	.57 (.04)
1 Day Delay	.45 (.03)	.53 (.03)
4 Day Delay	.44 (.03)	.50 (.04)

Mean (SE) performance during initial test blocks 1 and 2 for each retention interval group.

2.2.2. Final testing

2.2.2.1. Accuracy. The next analysis considers whether final recognition test accuracy and the magnitude of the testing effect differ as a function of retention interval. Accuracy was calculated as hits (confidence responses 4, 5, and 6 to old items) minus false alarms (FAs; confidence responses 4, 5, and 6 to new items) on the final recognition test.

Fig. 2 displays accuracy on the final recognition test by Initial Learning condition and retention interval group. As is apparent in the figure, performance was better for items previously tested than for items previously untested (i.e. a testing effect was revealed). Although overall performance declined following the immediate final test condition, the magnitude of the testing effect did not change across retention interval.

A Two-Way Mixed ANOVA examining the effect of retention interval (between-subjects: immediate, 1 day delay, 4 day delay) and Initial Learning condition (within-subjects: test, no test) on Final Testing accuracy supported this conclusion. There was a main effect of Initial Learning condition ($F(1, 80) = 155.04, p < .001, \eta_p^2 = .66$), revealing a testing effect in which performance was better for previously tested ($M = .39$) than for previously untested ($M = .27$) items.

There was, additionally, a main effect of retention interval ($F(2, 80) = 17.20, p < .001, \eta_p^2 = .30$). Post hoc comparisons with Tukey HSD correction for multiple comparisons revealed that performance was significantly better on the immediate final test ($M = .47$) than after both a 1 day delay ($M = .29$; mean difference = 0.18, corrected $p = .001$) and a 4 day delay ($M = .22$; mean difference = 0.25, corrected $p < .001$). Performance did not significantly differ between the 1 and 4 day retention interval conditions (mean difference = 0.07, corrected $p = .252$). Finally, the interaction between Initial Learning condition and retention interval was not significant ($F(2, 80) = 0.95, p = .392, \eta_p^2 = .02$), indicating that the magnitude of the testing effect did not change with increasing delay between Initial Learning and Final Testing.

For comparison, another common measure of accuracy, *d*’, was calculated in order to adjust for potential differences in response bias across subjects. Using this alternative measure of accuracy, all patterns of results remained the same with the exception that the interaction

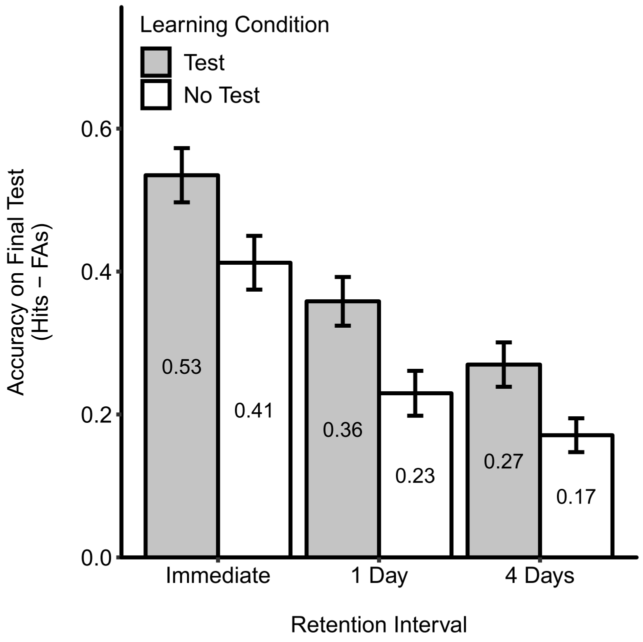


Fig. 2. Final recognition test accuracy in Experiment 1. Average accuracy (±SE) was calculated as correct recognition (hits) minus false alarms (FAs) and is depicted here by Initial Learning condition (test vs. no test) and retention interval (immediate, 1 day delay, 4 day delay).

between Initial Learning condition and retention interval approached significance ($F(2, 80) = 2.92$, $p = .060$, $\eta_p^2 = .07$).

2.2.2.2. Parameter estimates. The primary goal of Experiment 1 was to examine estimates of recollection and familiarity on the final test and the extent to which one or both of these processes support the testing effect seen at each delay. In order to obtain estimates of recollection and familiarity on the final test, confidence and accuracy data were fit to the Dual-Process Signal-Detection model using Maximum Likelihood Estimation (MLE) via the ROC Toolbox in MATLAB (Koen et al., 2017; Version 1.1.3). The ROC Toolbox creates individual subject ROCs and calculates recollection and familiarity parameter estimates, as well as subject-specific model fit statistics, for further analysis (see Koen et al., 2017; Yonelinas, 1994; Yonelinas et al., 2010; Yonelinas and Parks, 2007).

Model fit. We first address the fit of the Dual-Process Signal-Detection (DPSD) model to the confidence data. In each delay condition the average across subjects of individual-level R^2 goodness-of-fit measures was high: immediate final test mean (standard deviation) $R^2 = .96$ (0.04); 1-day delay $R^2 = .94$ (0.05); and 4-day delay $R^2 = .94$ (0.05). For the interested reader, we also fit the Experiment 1 confidence data to an alternative competing model, the Unequal-Variance Signal-Detection (UVSD) Model, which argues that a general memory strength signal, along with a variance parameter, can account for recognition memory data (e.g., see Wixted, 2007a). The UVSD model was likewise fit via the ROC Toolbox in MATLAB (Koen et al., 2017; Version 1.1.3). Both models provided a good fit to the data, with neither model unequivocally preferred, as is often the case (Parks and Yonelinas, 2007a). Supplementary Tables 1 and 2 provide more detailed information on the fit of the DPSD and UVSD models, for the interested reader.

The dual-process perspective and DPSD model, specifically, has been employed in a vast literature on various topics to estimate recollection and familiarity (e.g., see Yonelinas et al., 2010); and it is the perspective taken in the present study (see Diana et al., 2006; Jang et al., 2009, 2011; Parks and Yonelinas, 2007a, 2007b; Wixted, 2007a, 2007b; Wixted and Mickes, 2010; Yonelinas, 1994; Yonelinas and Parks, 2007, for debate on this subject). In the General Discussion, we provide a brief discussion of the debate between dual- and single- (and related) process views of recognition memory, along with an alternative interpretation of the present data in terms of the UVSD model.

Recollection and familiarity parameters. Fig. 3 displays parameter estimates of recollection (A) and familiarity (B) on the final recognition test by initial learning condition and retention interval group. As can be seen in Fig. 3A, estimates of recollection revealed a testing effect over time. Although recollection estimates declined following the immediate final test, the magnitude of the testing effect in recollection did not significantly change over time. Similarly, as can be seen in Fig. 3B, estimates of familiarity revealed a testing effect at all retention intervals. However, in addition to a decline in familiarity estimates following the immediate final test, the magnitude of the testing effect in familiarity likewise declined. Critically, however, a testing effect in familiarity estimates was revealed at all retention intervals.

A Three-Way Mixed ANOVA formally examined the effect of parameter (within-subjects: recollection, familiarity), Initial Learning condition (within-subjects: test, no test), and retention interval (between-subjects: immediate, 1 day delay, 4 day delay) on process estimates during Final Testing, revealing a significant three-way interaction ($F(2, 80) = 4.41$, $p = .015$, $\eta_p^2 = .10$). To break down this interaction, Two-Way Mixed ANOVAs were conducted separately for recollection and familiarity estimates to examine the effect of Initial Learning condition and retention interval on process estimates (see Supplementary Materials for the results of Two-Way Repeated Measures ANOVAs for each retention interval).

For estimates of recollection, the main effect of Initial Learning condition was significant ($F(1, 80) = 80.20$, $p < .001$, $\eta_p^2 = .50$; $M_{\text{test}} = .24$, $M_{\text{no test}} = .12$), revealing a testing effect in recollection estimates on the final test. There was also a significant main effect of retention interval ($F(2, 80) = 10.68$, $p < .001$, $\eta_p^2 = .21$); post hoc comparisons with Tukey HSD correction revealed that performance was significantly better on the immediate final test ($M = .27$) than after both a 1 day delay ($M = .17$; mean difference = 0.10, corrected $p = .024$) and a 4 day delay ($M = .10$; mean difference = 0.17, corrected $p < .001$). Performance did not significantly differ between the 1 and 4 day retention interval conditions (mean difference = 0.07, corrected $p = .198$). Finally, the two-way interaction between initial learning condition and retention interval was not significant ($F(2, 80) = 1.51$, $p = .227$, $\eta_p^2 = .04$), indicating

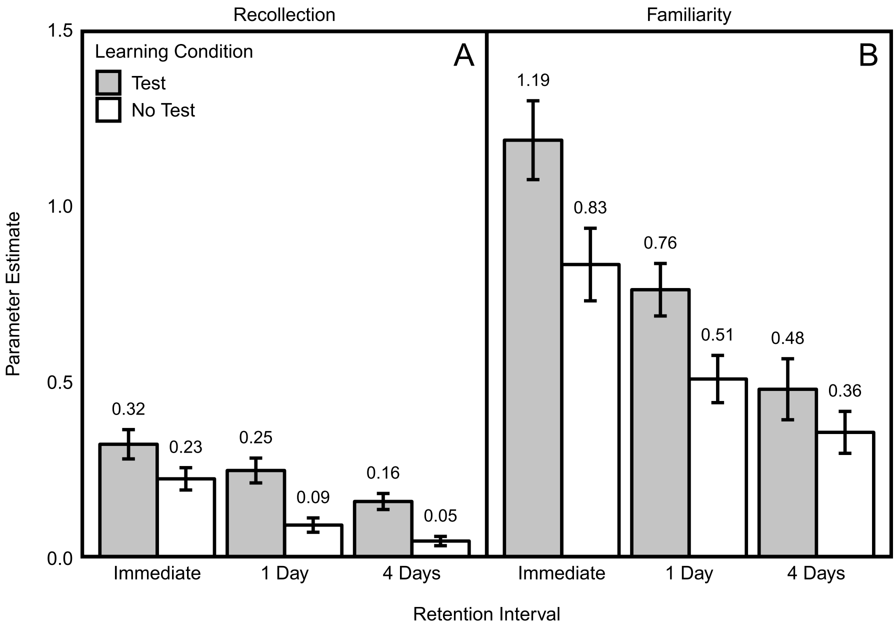


Fig. 3. Parameter estimates on the final recognition test in Experiment 1. Average estimates (±SE) of (**A**) recollection and (**B**) familiarity on the final recognition test are depicted by Initial Learning condition (test vs. no test) and retention interval (immediate, 1 day delay, 4 days delay).

that the magnitude of the testing effect in recollection estimates did not significantly change with increasing delay between Initial Learning and Final Testing.

For estimates of familiarity, by contrast, the two-way interaction between Initial Learning condition and retention interval was significant ($F(2, 80) = 5.05$, $p = .009$, $\eta_p^2 = .11$), indicating that the magnitude of the testing effect in familiarity estimates changed with increasing retention intervals between Initial Learning and Final Testing. To explore this interaction further, post hoc analyses examined the effect of Initial Learning condition on estimates of familiarity at each delay, finding a significant testing effect on an immediate ($t(28) = 6.55$, $p < .001$, Cohen's $d = 1.22$), 1-day delayed ($t(24) = 5.07$, $p < .001$, Cohen's $d = 1.01$), and 4-day delayed ($t(28) = 2.31$, $p = .029$, Cohen's $d = .43$) final test.

Post hoc analyses of the effect of retention interval on estimates of familiarity in both the test and no test condition were likewise examined, finding a significant effect of delay in the test condition ($F(2, 80) = 14.99$, $p < .001$, $\eta_p^2 = .27$). Follow-up pairwise comparisons with Tukey HSD correction revealed that estimates of familiarity in the testing condition were significantly higher on the immediate final test ($M = 1.19$) than after both a 1-day delay ($M = .76$; mean difference = 0.42, corrected $p = .007$) and a 4-day delay ($M = .48$; mean difference = 0.71, corrected $p < .001$). However, estimates of familiarity in the testing condition did not significantly differ between the 1- and 4-day retention interval conditions (mean difference = 0.28, corrected $p = .098$). There was likewise a significant effect of delay in the no test condition ($F(2, 80) = 9.65$, $p < .001$, $\eta_p^2 = .19$); pairwise comparisons with Tukey HSD correction revealed that estimates of familiarity in the no testing condition were significantly higher on the immediate final test ($M = 0.83$) than after both a 1-day delay ($M = .51$; mean difference = 0.33, corrected $p = .016$) and a 4-day delay ($M = .36$; mean difference = 0.48, corrected $p < .001$). However, as before, estimates of familiarity in the no testing condition did not significantly differ between the 1- and 4-day retention interval conditions (mean difference = 0.15, corrected $p = .390$).

In summary, although some differences in the pattern of results for recollection and familiarity estimates were observed, a testing effect was revealed in estimates of recollection and familiarity at all retention intervals.

2.3. Discussion

The central finding of Experiment 1 was that prior retrieval practice magnified estimates not just of recollection but also of familiarity on a later final test. This pattern of results occurred when the final test immediately followed initial studying and testing, and it was maintained across a 1- and 4-day retention interval between initial and final testing. Thus, in Experiment 1 we did not replicate prior literature that posits a role for recollection only in producing the testing effect on an immediate final test. Of note, the method of estimating recollection and familiarity in the current study (confidence responses fit to the Dual-Process Signal Detection model) aligns with the method used on delayed final tests in Bies-Hernandez (2013) but differs from the methods used on immediate final tests in prior studies. Experiment 2 was designed to address whether this methodological discrepancy might account for the present findings and dictate whether a testing effect in familiarity would be observed on an immediate final test. As such, in Experiment 2 another common method of estimation (and one used in Chan and McDermott, 2007) was adopted: the Remember-Know technique (Tulving, 1985; Yonelinas, 2002; Yonelinas and Jacoby, 1995).

3. Experiment 2: Conceptual replication

The purpose of this study was to establish the generalizability of the conclusions of Experiment 1 with a new measure for estimating recollection and familiarity (one used in prior research with an immediate final test). Performance was examined on an immediate and 4-day delayed final test, using Remember-Know-New responses from which to estimate recollection and familiarity. A 1-day delayed final test was not included, as the 1- and 4-day delayed final tests produced virtually identical patterns of results in Experiment 1.

3.1. Materials and methods

3.1.1. Participants

One-hundred and twenty-nine subjects completed an initial training session of the experiment on Amazon Mechanical Turk (immediate final test condition: 59; 4-day delay: 70). Of these, 7 subjects were excluded because they failed to complete the remaining experimental session(s) (immediate final test condition: 1; 4-day delay: 6). Specifically, in order to ensure that subjects understood the Remember-Know-New (RKN) instructions, all subjects completed an extensive training session before being admitted to the main experiment (see Procedure section below for details regarding the training). An experimenter then judged responses from the training task and determined whether the subject understood the RKN distinction (responses were judged twice by the same experimenter at two different time points; any disagreements between the judgments were considered and then resolved). Participants who did not pass the RKN training were not invited to complete the main experiment. Forty-five subjects failed to pass the RKN training (immediate final test: 22; 4-day delay: 23; see Supplementary Materials for example correct and incorrect responses). An additional 38 subjects were excluded due to a combination of the following issues (for some subjects, the following reasons for exclusion occurred in conjunction with failing the initial RKN training or failing to complete all remaining experimental sessions]): noting words during the task (8), restarting the experiment after beginning the main task (1), missing data (2), reporting not normal or corrected-to-normal vision (3), reporting a possible neurological disorder (6), reporting that English is not the subject's native language (1), reporting that the subject does not wish to be invited back to complete the main experimental task (1), and failing a short RKN retraining before completing the final recognition test (21; this retraining was completed to ensure that subjects continued to exhibit an adequate understanding of the RKN distinctions before completing the Final Testing session). One additional subject was excluded as final test responses clearly indicated a complete lack of engagement in the task (making over 50 of the same responses in a row). All subjects were at least 18 years old. Forty-three subjects met all inclusion criteria: immediate final test (N = 21, mean age = 34.8, SD age = 9.0, age range = 23–51, female [F] = 8), 4-day delay (N = 22, mean age = 37.0, SD age = 9.1, age range = 27–63, F = 11). Conditions were run sequentially on Amazon Mechanical Turk (1: immediate; 2: 4-day delay).

Subjects were compensated approximately $9.00 per hour for their participation (and an additional $1.00 for completing the initial RKN training). Before beginning the study, subjects consented to participation and, upon completion of the study, were debriefed. Experiment 2 was conducted in accordance with the Washington University in St. Louis Institutional Review Board.

3.1.2. Materials

All experimental task materials were the same as in Experiment 1, with one exception. Specifically, one stimulus in the current study, "REMEMBER," was replaced with another stimulus of the same length and Kučera-Francis Frequency during data collection in the immediate final test condition when it came to the experimenter's attention that the stimulus may be confusing for subjects (given that the final test response instructions for Experiment 2 contain the response choice "Remember"; see below). Responses to the word "REMEMBER" were removed from analysis for any subjects in Experiment 2 who completed the task before the stimulus was replaced.

RKN training and retraining materials consisted of explanations of the RKN response options, 12 multiple choice questions, and 6 short answer questions concerning the RKN judgments and distinctions. The complete RKN training task can be found in the Supplementary Materials along with example responses from one subject who passed and one subject who failed the training.

3.1.3. Procedure

3.1.3.1. Initial RKN training. To ensure that subjects understood the RKN distinctions and task instructions—and because misunderstandings create difficulty in interpretation of the recollection and familiarity estimates obtained from RKN responses—before completing the main experimental task, subjects completed an RKN training session (see RKN Training in the Supplementary Materials for the complete RKN training instruction). During the training, subjects were provided with explanations of each response type and were asked to review these instructions twice. Subjects then answered 12 multiple choice questions concerning the response types (4 questions per response type; e.g. “You are presented with the word ‘FRAME.’ You recall that this had come right at the beginning of the previous list. What response do you make?” With answer choices: Remember, Know, New). Subjects were then asked to provide a description and an example (in separate short answer forms) of each of the three judgment types. On the basis of responses to the multiple choice and short answer questions, an experimenter determined whether the subject understood the RKN distinction and instructions. Subjects who understood the RKN distinction were invited back the next day to take part in the main experimental task.

3.1.3.2. Initial learning. The Initial Learning procedure was identical to that of Experiment 1.

3.1.3.3. Retention interval. Subjects completed the final recognition test either 1) immediately (average retention interval = 7.54 min [longer than the equivalent immediate final test condition in Experiment 1 due to the addition of an RKN retraining before the final test, see RKN retraining below], SD = 3.48 min, range = 3.27 min–15.00 min); or 2) approximately 4 days later (average retention interval = 4.07 days, SD = 0.22 days, range = 3.94 days–4.93 days).

3.1.3.4. RKN retraining. Before the final recognition test, subjects completed an RKN retraining task. The RKN retraining task was similar to the initial RKN training task, except that it did not include a multiple choice section. All subjects who completed the RKN retraining had the option to continue to the final recognition test following the retraining (the final recognition test followed immediately and automatically after the RKN retraining). However, an experimenter reviewed RKN retraining responses after-the-fact, and excluded from analysis any subjects that did not appear to understand of the RKN distinction and instructions during the retraining (see the Participants section above).

3.1.3.5. Final testing. The final recognition test was identical to that of Experiment 1, with the exception of the response options and instructions. Subjects were asked to indicate their memory for stimuli by making Remember-Know-New responses via the 7-8-9 keys (or 9-8-7 keys, for a portion of the subjects) on their keyboard. A Remember response indicates that the participants can consciously recall specific parts of the experience they had when they saw the word in the Initial Learning session of the study. By contrast, a Know response indicates that the participants have a gut feeling that they saw the word in the previous session of this study, but that they do not have conscious recollection of ‘seeing’ or ‘experiencing’ it in the previous session. A New response indicates that the participant believes the word to be new to the study.

3.2. Results

As in Experiment 1, the existence of outliers (more than 3 standard deviations from the mean) was examined for each analysis. No outliers were detected in Experiment 2.

3.2.1. Initial learning

First, initial learning was examined to ensure that performance was similar across the different retention interval groups. Accuracy was calculated for each subject in the same way as in Experiment 1.

As in Experiment 1, initial test accuracy was similar in the immediate and 4 day delay retention interval groups (see Table 2). A Two-Way Mixed ANOVA examined the effect of retention interval condition (between-subjects: immediate, 4 day delay) and test block (within-subjects: Initial Test Block 1, Initial Test Block 2) on initial test accuracy. The main effect of retention interval was not significant ($F(1, 41) = 0.01, p = .917, \eta_p^2 < 0.001$); subjects in both retention interval conditions performed similarly on initial tests ($M_{\text{immediate}} = .49$, $M_{4\ \text{day}} = .49$). In addition, there was no significant main effect of initial test block, although subjects numerically improved from the first ($M = .47$) to the second ($M = .51$) initial cued-recall test ($F(1, 41) = 3.91, p = .055, \eta_p^2 = .09$). Finally, the interaction between retention interval condition and initial test number was not significant ($F(1, 41) = 0.31, p = .578, \eta_p^2 = .01$). Thus, performance on initial tests was roughly equivalent across retention interval groups and reflected a pattern of results similar to that found in Experiment 1.

3.2.2. Final testing

3.2.2.1. Accuracy. Accuracy on the recognition test was calculated as hits (Remember and Know responses to items previously seen in the experiment) minus FAs (Remember and Know responses to items not previously seen in the experiment).

Fig. 4 displays accuracy on the final recognition test by Initial Learning condition and retention interval group. As in Experiment 1, a testing effect occurred across retention interval. Although overall performance declined over time, the magnitude of the testing effect did not change over time.

A Two-Way Mixed ANOVA examined the effect of retention interval group (between-subjects: immediate, 4 day delay) and Initial Learning condition (within-subjects: test, no test) on Final Testing accuracy. There was a main effect of Initial Learning condition ($F(1, 41) = 76.40, p < .001, \eta_p^2 = .65$); a testing effect emerged such that performance was greater for items previously tested ($M = .46$) than for items previously untested ($M = .33$). There was, additionally, a main effect of retention interval ($F(1, 41) = 35.96, p < .001, \eta_p^2 = .47$); as expected, performance was better on the immediate final test ($M = .52$) than after a 4-day delay ($M = .27$). As in Experiment 1, the interaction between Initial Learning condition and retention interval was not significant ($F(1, 41) = 0.70, p = .407, \eta_p^2 = .02$), indicating that the magnitude of the testing effect did not differ as a function of retention interval between Initial Learning and Final Testing.

For comparison, another common measure of accuracy, *d'*, was calculated to adjust for potential differences in response bias across

Table 2
Initial cued-recall test performance in Experiment 2. Means (standard errors) are reported for initial test blocks 1 and 2 by retention interval group (immediate, 4 day delayed).

Experiment 2: Initial Cued-Recall Test Accuracy		
Retention Interval	Initial Test Block 1	Initial Test Block 2
Immediate	.47 (.03)	.50 (.03)
4 Day Delay	.46 (.03)	.52 (.04)

Mean (SE) performance during initial test blocks 1 and 2 for each retention interval group.

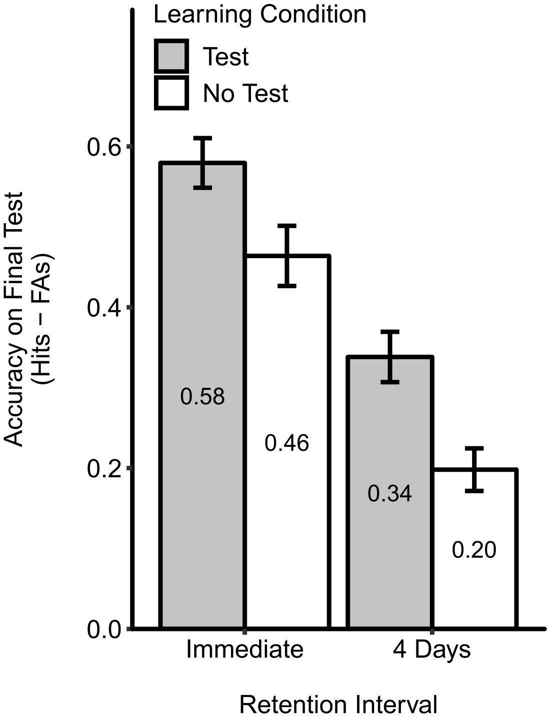


Fig. 4. Final recognition test accuracy in Experiment 2. Final recognition test average accuracy (±SE) was calculated as correct recognition (Remember and Know hits) minus false alarms (Remember and Know FAs) and is depicted here by Initial Learning condition (test vs. no test) and retention interval (immediate, 4 day delay).

subjects. Using this alternative measure of accuracy, all patterns of results remained the same.

3.2.2.2. Parameter estimates. The primary goal in conducting Experiment 2 was to examine the extent to which recollection and familiarity supported the testing effect over time, using RKN data in order to obtain process estimates. Estimates of recollection and familiarity on the final recognition test were calculated by using the Independence Remember-Know Procedure (Yonelinas, 2002; Yonelinas and Jacoby, 1995; see Supplementary Materials for exact formulas used).

Fig. 5 displays parameter estimates of recollection (A) and familiarity (B) on the final recognition test by Initial Learning condition and retention interval group. As can be seen in the figure, estimates of both recollection and familiarity revealed a testing effect over time. Although overall recollection and familiarity estimates declined from the immediate to the 4-day delayed final test, the magnitude of the testing effect in recollection and familiarity did not significantly change over time.

A Three-Way Mixed ANOVA formally examined the effect of parameter (within-subjects: recollection, familiarity), Initial Learning condition (within-subjects: test, no test), and retention interval (between-subjects: immediate, 4 day delay) on process estimates during Final Testing. The three-way interaction was not significant ($F(1, 41) = 2.12, p = .153, \eta_p^2 = .05$). In addition, no significant two-way interactions were present (parameter X retention interval, $p = .951$; Initial Learning condition X retention interval, $p = .567$; parameter X Initial Learning condition, $p = .847$).

Lastly, main effects of Initial Learning condition, retention interval, and parameter were examined. The main effect of Initial Learning condition was significant ($F(1, 41) = 75.93, p < .001, \eta_p^2 = .65$), indicating a testing effect, such that process estimates were higher for items previously tested ($M = .34$) than for items previously untested ($M = .23$). The main effect of retention interval was also significant ($F(1, 41) = 29.54, p < .001, \eta_p^2 = .42$), indicating that process estimates were higher for items on an immediate final test ($M = .38$) than for items on a 4-day delayed final test ($M = .19$). The main effect of parameter, however, was not significant ($F(1, 41) = 1.92, p = .173, \eta_p^2 = .04$).

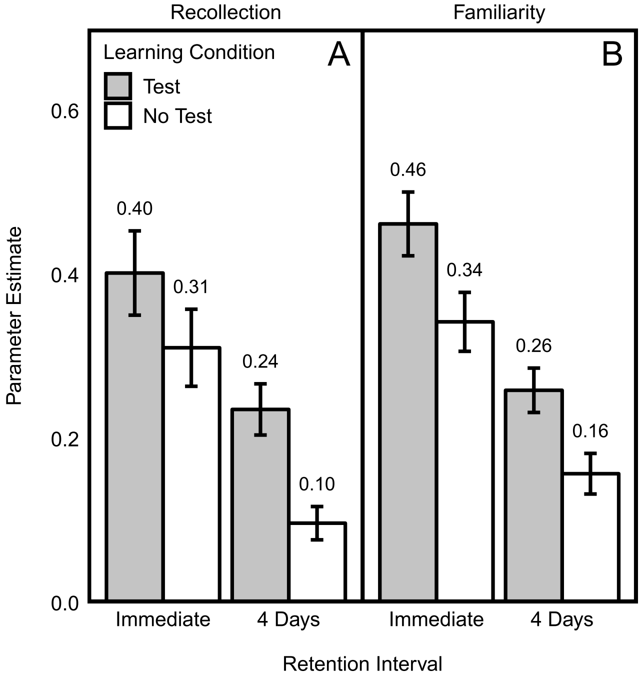


Fig. 5. Parameter estimates on the final recognition test in Experiment 2. Average estimates (±SE) of (**A**) recollection and (**B**) familiarity on the final recognition test are depicted by Initial Learning condition (test vs. no test) and retention interval (immediate, 4 day delay).

Thus, as in Experiment 1, at both retention intervals, estimates of both recollection and familiarity revealed a testing effect during Final Testing, and process estimates decreased over time. In contrast to Experiment 1, the magnitude of the testing effect in recollection and familiarity did not differentially change over time.

3.2.2.3. Raw know responses. Of note, although estimates of familiarity revealed a clear testing effect at both retention intervals, raw Know response probabilities did not (immediate: $t(20) = 1.00$, $p = .329$, Cohen's $d = 0.22$; 4-day delay: $t(21) = 0.08, p = .941$, Cohen's $d = 0.02$; see Supplementary Fig. 1). Although an overall testing effect in Know responses across subjects was not observed, some subjects did exhibit a numerical testing effect in Know responses. The lack of an overall testing effect in raw Know probabilities is likely due to the fact that, to the extent that prior testing enhances recollection, fewer Know responses can be made. This pattern occurs because in the Remember-Know procedure, when a participant recollects an item, he or she must provide a Remember, rather than a Know, response, even if item familiarity was experienced. The Independence Remember-Know Procedure, commonly used to obtain parameter estimates of familiarity (and used in the present study), serves to adjust for this bias.

Beyond this suggestion, in the present data we observe strong evidence to suggest that the lack of an observed testing effect in raw Know probabilities is not due to a lack of an effect of prior testing on familiarity. Instead, evidence suggests that enhanced recollection following testing led to fewer Know responses for items previously tested, even when familiarity was enhanced due to prior testing. Specifically, the magnitude of the testing effect in raw Know responses revealed a moderate-to-strong negative correlation with the magnitude of the testing effect in raw Remember responses (immediate: $r(19) = -.64, p = .002$; 4-day delay: $r_s(20) = -.44$, $p = .040$; see Supplementary Fig. 2).

This negative relation suggests one of two possibilities. One possibility, which we believe to be very unlikely, is that when prior testing enhances recollection more for an individual, familiarity is truly enhanced less for that individual. The second possibility, which we believe to be much more likely, is that individuals for whom prior testing enhances recollection more simply cannot provide Know responses as readily to items previously tested (given their enhanced Remember responding). This situation then produces the negative relation across subjects between the magnitude of the testing effect in raw Remember and raw Know responses. Thus, the absence of an overall testing effect in raw Know probabilities does not pose issues for the above conclusions with regard to estimates of familiarity.

The primary goal of Exp. 2 was to examine parameter estimates of recollection and familiarity. As such, the analysis of raw Know probabilities was conducted for the purpose of completeness and will not be considered further.

3.3. Discussion

Experiment 2 was designed to examine the generalizability of the finding that prior testing magnified estimates of both recollection and familiarity on an immediate and a delayed final test. Critically, Experiment 2 used a different method of obtaining estimates of recollection and familiarity (the Remember-Know procedure, used in Chan and McDermott, 2007), and obtained the same pattern of results: prior retrieval practice improved both recollection and familiarity processing on an immediate and 4-day delayed final test.

4. General Discussion

The present study examined the mechanisms that support the benefits of retrieval practice to later memory. Of particular interest were two processes—recollection and familiarity—initially proposed by Endel Tulving (1985) to characterize distinct memory systems (also see Jacoby, 1991; Yonelinas, 2002). Since its inception, the recollection-familiarity distinction has been used to understand a large variety of memory phenomena. In the present study, we used this distinction to better understand a robust effect in the memory literature: the testing effect.

Prior research using short retention intervals of approximately 15 min or less between initial and final testing has found that retrieval practice magnifies estimates of recollection but does not affect the use of familiarity (Chan and McDermott, 2007; Pu and Tse, 2014; Verkoeijen et al., 2011; although see Gao et al., 2016). However, several findings suggest that the mechanisms that support the long-term testing effect, one obtained with retention intervals of a day or longer, may differ from those previously found to support the effect on an immediate final test. This suggestion has come from research that observes change over time in the magnitude and direction of the testing effect (Roediger and Karpicke, 2006a), as well as change over time and across tests in the use of recollection and familiarity (Conway et al., 1997; Dewhurst et al., 2009; Gardiner and Java, 1991). This suggestion has likewise come from studies in which delayed estimates of familiarity following initial testing reveal a testing effect. However, with the exception of one study (Bies-Hernandez, 2013), these estimates were indirectly assessed by the present authors via group means reported in the papers (Dudukovic et al., 2009; Kessler et al., 2014).

On an immediate, 1-day delayed (Exp. 1 only), and 4-day delayed final test, the present study observed a testing effect in estimates of both recollection and familiarity. This pattern was found when measures of recollection and familiarity were obtained via confidence (Exp. 1) and Remember-Know-New (Exp. 2) judgments. Thus, contrary to the majority of prior work using short retention intervals, the present results suggest that the benefits of retrieval practice can arise from changes in both familiarity and recollection. In considering only a role for recollective processes in supporting the testing effect, current explanations of the testing effect may be incomplete. Below, the present results are considered in light of prior research, and we provide a working hypothesis as to some of the key variables that may impact whether a familiarity-based contribution to the testing effect will emerge on a final test. We hope this discussion will shed light on potential explanations for the discrepancies in the literature, as well as provide testable predictions as to when a testing effect in familiarity estimates may be revealed on a final test.

4.1. The effect of delay and other design characteristics

Prior to the current study, only a handful of studies have used the dual-process perspective to directly examine the mechanisms that support the benefit of retrieval practice to later memory. Whereas the majority of these studies employed a short retention interval between initial and final testing and found increases only in estimates of recollection following initial testing (Chan and McDermott, 2007; Pu and Tse, 2014; Rowland, 2011; and Verkoeijen et al., 2011; but see Gao et al., 2016), one study employed a long retention interval and found increases in familiarity following initial testing, and a less reliable increase in recollection (Bies-Hernandez, 2013). The current study used an immediate, 1-day delayed, and 4-day delayed final test to explore the impact of retention interval, and, in all cases, found reliable increases in both familiarity and recollection following initial testing. In light of these results, one might conclude that retention interval, per se, is not a key factor in determining whether a testing effect in familarity estimates is revealed on a final test. However, it may be premature to draw this conclusion. We propose that, although delay may not be *necessary* in order to observe a benefit of testing to familiarity, it may be *sufficient*.

Indeed, inspection of the current experiments, along with those discussed in the Introduction to this article, reveals that, for any study in which the retention interval between initial studying and final testing is 1 day or longer, a testing effect in familiarity-related processes is observed either statistically or numerically. This outcome was observed with a 1-day retention interval (Kessler et al., 2014, and the present study Exp. 1); a 2-day retention interval (Bies-Hernandez, 2013, Exp. 1 and Exp. 2 and Dudukovic et al., 2009, Exp. 1); and a 4-day retention interval (the present study Exp. 1 and Exp. 2). By contrast, each study in which a testing effect in familiarity estimates is not observed used very short retention intervals (Chan and McDermott, 2007, Exps. 1–3; Verkoeijen et al., 2011, Exps. 1–4; Pu and Tse, 2014; Rowland, 2011, Exp. 1; and Jones and Roediger, 1995). Why then, in the present study is a testing effect in familiarity estimates on an immediate final test reliably observed? One possibility is that, although retention interval may be an important factor, several other design characteristics interact to reveal or fail to reveal a benefit of testing to familiarity.

Before proposing several potentially influential characteristics, we seek to address why, theoretically, if prior testing truly enhances familiarity processing, might increasing the retention interval between initial learning and final testing increase the likelihood of observing a testing effect in familiarity estimates. One possibility is that prior testing only *directly* enhances recollection, and, by contrast, *indirectly* enhances familiarity. The suggestion is that, over time, what was initially enhanced recollection due to prior retrieval practice becomes decontextualized memory, and is, thus, revealed on a delayed final test as enhanced familiarity. This suggestion would be in line with findings from Dewhurst et al. (2009) and Conway et al. (1997), in which the authors observed that, with repeated testing, many correct responses initially accompanied by conscious recollection over longer retention intervals become correct responses unaccompanied by conscious recollection.

Although an *indirect* benefit of testing to familiarity could account for the fact that a testing effect in familiarity is more often observed after longer retention intervals, this explanation cannot account for the current result: that testing can, in some cases, magnify estimates of familiarity even on an immediate final test. Another possibility is that, although retrieval practice may *directly* enhance familiarity, this benefit to familiarity will often fail to appear on an immediate final test. Specifically, on an immediate final test, when recollection is presumably fairly high, subjects may rely successfully on recollection processes in order to succeed in the task (see Bies-Hernandez, 2013, and Chan and McDermott, 2007, for related arguments with respect to reliance on recollective processes during final testing). When a longer retention interval is introduced, however, and recollection of the initial learning experience presumably decreases, subjects may rely relatively less on recollective processing for success in the task. In this circumstance, a testing effect in familiarity—which may always have existed, even if unobserved—will more consistently appear on the final test. This possibility would suggest that in an experiment that employs an immediate final test, various design choices at any stage of the experiment that emphasize, encourage, or improve recollective processing would lower the likelihood of observing a testing effect in familiarity estimates. Conversely, elements of an experiment that place comparatively less emphasis on the use of recollective processing or make recollective processing more difficult should increase the likelihood of observing a testing effect in familiarity estimates on an immediate final test.

Prior work has considered this possibility with respect to final test format (Bies-Hernandez, 2013). Indeed, whereas source memory and exclusion tests emphasize the use of recollection to succeed on the final test, confidence-based and Remember-Know judgments place no such emphasis. Thus, a testing effect in familiarity may be more readily observed when using confidence and Remember-Know final test formats because subjects need not attend specifically to recollective processes in order to successfully complete the task (Bies-Hernandez, 2013; also see Chan and McDermott, 2007, for a related argument). In line with this suggestion, we observed that, of the studies that found a negative or non-existent effect of testing on estimates of familiarity (Chan and McDermott, 2007, Exps. 1–3; Verkoeijen et al., 2011, Exps. 1–4; Pu and Tse, 2014; Rowland, 2011, Exp. 1; and Jones and Roediger, 1995), more than half used a source memory or exclusion final test. By contrast, of the studies that revealed a positive statistical or numerical testing effect in familiarity-related processes (the current study, Exp. 1 and Exp. 2; Bies-Hernandez, 2013, Exp. 1 and Exp. 2; Dudukovic et al., 2009, Exp. 1; Gao et al., 2016; and Kessler et al., 2014), over half used a Remember-Know or confidence-based final test.

In addition to final test emphasis on recollection, we suggest that design choices *at any stage* of the experiment that emphasize or improve recollective processing may lower the likelihood of observing a testing effect in familiarity estimates on an *immediate* final test. In line with this suggestion, we observed that the number of items to be remembered in the testing condition was lower in studies that have failed to find familiarity-based contributions to the testing effect (30–64 items) than in studies that have found a testing effect in familiarity (40–140 items). Perhaps with fewer items to remember on an immediate final test, recollection is relatively high and can be more readily relied on for success in the task. However, as the number of items to be remembered increases, recollection may be more heavily taxed. As a result, participants may be less able to rely upon recollection for success in the task.

In the present data it is possible to explore this suggestion to some extent through an examination of individual differences in overall recollection and its relation to the magnitude of the testing effect in familiarity. In other words, across subjects does higher recollection predict reductions in the testing effect in familiarity? When examining parameter estimates of recollection and familiarity, there was no consistent statistical relation between the average of recollection estimates in the test and no test condition and the magnitude of the testing effect in familiarity (correlation coefficients and significance tests for these analyses are provided in Supplementary Table 3).

By contrast, examination of raw confidence (Exp. 1) and Remember-Know (Exp. 2) responses provided a different picture. Specifically, for the raw confidence (Exp. 1) and Remember-Know (Exp. 2) data, average raw recollection-related responding was calculated for each subject as the average proportion of “6” (highest confidence; Exp. 1) or “Remember” (Exp. 2) hits minus false alarms in the test and no test conditions. The magnitude of the testing effect in familiarity-related responding was calculated as the proportion of “4” and “5” (lower confidence; Exp. 1) or “Know” (Exp. 2) hits in the test condition minus that in the no test condition. Using raw confidence and Remember-Know responses, a negative relation was revealed between average recollection-related responding and the magnitude of the testing effect in familiarity-related responding on the immediate and 1-day delayed final tests (Exp. 1 only). However, this relation was not observed statistically at the 4-day delay (although the relationships remained negative in direction; see Supplementary Table 3).

Thus, across the literature it appears that the magnitude of the testing effect in familiarity may be influenced in part by the extent to which the study emphasizes recollection throughout. However, in the present study, the relation between an individual’s average recollection and the magnitude of the testing effect in familiarity was not clear-cut.

Finally, we observed that experiments that did not obtain a testing effect in overall accuracy and that included a restudy, rather than no test, control condition less often observed a testing effect in familiarity estimates.

Of note, the above possibilities of indirect and direct effects of testing on familiarity are not mutually exclusive and might, together, help to explain why it is less likely to observe a testing effect in familiarity on an immediate final test and more likely on a delayed final test. Further, the possibilities discussed above are only speculative; experimental examination will be necessary before drawing strong conclusions.

4.2. *Alternative interpretations*

This article and the literature it is founded on functions within the dual-process perspective of recognition memory. This perspective, and the idea that independent recollection and familiarity processes can be estimated from recognition memory performance, has been critically examined in the literature, with much evidence to suggest its validity, and the validity of the Dual-Process Signal-Detection (DPSD) model, specifically (e.g., see Parks and Yonelinas, 2007a; Yonelinas, 2001, 2002).

However, a competing perspective suggests that a general memory strength signal can better account for recognition memory performance (e.g., see Wixted, 2007a, 2007b). A prominent model from this perspective, the Unequal-Variance Signal-Detection (UVSD) model, argues that a single memory strength signal, along with unequal variances of old item (previously presented) and new item (not previously presented) memory strength distributions, can account for recognition performance (Wixted, 2007a). Critically, in this framework, whether or not recollection or familiarity processes underlie the single memory strength signal, the two processes cannot be separately estimated (Wixted, 2007a, 2007b).

Interpretation of the present results within this alternative framework would suggest that prior testing broadly increases the memory strength signal for tested items, leading to improved recognition performance (i.e. the testing effect). In the present study, Experiment 1 confidence data were fit to both the DPSD and UVSD models, for comparison. Results indicated that both models fit the data well (and to a similar degree), as is often the case (see Supplementary Tables 1–2).

Critically, in the present study, recollection and familiarity were additionally estimated via Remember-Know responses (Experiment 2). An important consideration in interpretation of Remember-Know responses is the extent to which subjects correctly interpret and apply the Remember-Know distinction when responding on the final test (e.g., for discussion see Migo et al., 2012; Parks and Yonelinas, 2007a; McCabe and Geraci, 2009; but also see Rotello et al., 2005). Briefly, it is important to ensure that Remember and Know responses indicate qualitative differences in memory and do not merely indicate responding in terms of somewhat higher vs. somewhat lower memory strength, respectively (also see Wixted, 2007a, 2007b; Parks and Yonelinas, 2007a). To the extent that Remember and Know responses do simply indicate somewhat higher vs. somewhat lower memory strength, interpretation of the present results from a single-process (or related) perspective would suggest that prior testing leads to an increase in the proportion of stronger memories (raw Remember responses) and no change in the proportion of weaker memories (raw Know responses). Thus, in order to make sense of these data from the dual-process perspective of recognition memory, it is important to verify that subjects correctly interpret and apply the Remember-Know distinction when responding on the final test (e.g., Migo et al., 2012).

To this end, prior to Experiment 2 participants completed an extensive Remember-Know training session, and only subjects who demonstrated clear understanding of the distinction were invited to take part in the main experimental task. Participants further completed a Remember-Know re-training session immediately preceding the final Remember-Know recognition test. These steps were undertaken to help ensure that only subjects who were correctly interpreting the Remember-Know instructions during the final test were included in analysis. The large degree of alignment between the results of Experiment 1 and 2, when using completely different methods for estimating recollection and familiarity, along with the extensive Remember-Know training, provides increased confidence in the conclusions drawn in the present study.

The goal of the present article is to engage in debate with the host of prior literature which suggests that the benefits of retrieval practice function only through recollection-related processes (if one subscribes to the dual-process perspective). Thus it is beyond the scope of this article to provide a comprehensive review of the debate between dual- and single-process (or related) perspectives of recognition memory. However, we believe this debate to be of great importance and direct the interested reader to a thorough and continuing discussion on the topic (e.g. see Diana et al., 2006; Dunn, 2004; Jang et al., 2009, 2011; Parks and Yonelinas, 2007a, 2007b; Rotello et al., 2006; Wixted, 2007a, 2007b; Wixted and Mickes, 2010; Yonelinas, 1994; Yonelinas and Parks, 2007).

4.3. Implications: Neural correlates of the testing effect over time

There is little agreement as to the neural mechanisms that support the testing effect (van den Broek et al., 2016). In light of prior research examining the neural correlates of recollection and familiarity during recognition testing, however, the present results point to several suggestions as to potential neural correlates of the testing effect in recognition memory. To the extent that recollection-related processes during recognition memory are supported by regions in the inferior lateral parietal cortex, hippocampus, and PhC (Diana et al., 2013; Ranganath, 2010; Vilberg and Rugg, 2007; Yonelinas et al., 2005), one could expect that these regions may be involved in producing the testing effect on both immediate and delayed final tests. Similarly, to the extent that familiarity-related processes are supported by regions in the superior lateral parietal cortex and PrC (Ranganath, 2010; Vilberg and Rugg, 2007; but see Diana et al., 2013), one could predict that these regions may likewise be involved in producing both the immediate and delayed testing effect. Future studies that compare retrieval-related activation during recognition testing for items previously tested and items previously untested on both immediate and delayed final tests could help to address this question and improve our understanding of the neural mechanisms that support the testing effect.

4.4. Conclusions

The benefit of testing may, in many cases, accrue via enhanced recollective processes. However, the present results clearly demonstrate that both the short- and long-term benefits of testing can, in some cases, be due to changes in both recollection and familiarity processes. Thus, the answer as to whether recollection or familiarity processes support the testing effect may simply be: it depends. Future work that seeks to examine the variables that influence the extent to which familiarity is revealed as a supporting mechanism on both immediate and delayed final tests will advance our understanding of the effect. If familiarity processes reliably support the testing effect under particular conditions, theoretical explanations of the effect would be improved by accounting for a role for familiarity.

Funding

This work was supported by a Collaborative Activity Award from the James S. McDonnell Foundation (Applying Cognitive Psychology to Enhance Educational Practice).

Author contributions

Ruth Shaffer and Kathleen McDermott conceptualized and designed the studies. Ruth Shaffer collected data, performed data analysis, and wrote the original draft of the manuscript. Both authors contributed to manuscript revisions.

Declaration of competing interest

None.

Acknowledgements

This material is based upon work supported by the National Science Foundation Graduate Research Fellowship Program under Grant No. DGE-1745038 to R.A.S. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation. We are grateful to Adrian Gilmore, Andy DeSoto, Jason Chan, and Roddy Roediger for helpful discussion on this project.

Data: All data can be accessed via the Open Science Framework https://osf.io/vhaxy/.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.neuropsychologia.2019.107298.

References

Anderson, N.D., Ebert, P.L., Jennings, J.M., Grady, C.L., Cabeza, R., Graham, S.J., 2008. Recollection- and familiarity-based memory in healthy aging and amnestic mild cognitive impairment. Neuropsychology 22 (2), 177–187. https://doi.org/10.1037/0894-4105.22.2.177.

Balota, D.A., Yap, M.J., Cortese, M.J., Hutchison, K.A., Kessler, B., Loftis, B., et al., 2007. The English Lexicon project. Behav. Res. Methods 39 (3), 445–459. https://doi.org/10.3758/BF03193014.

Bastin, C., Van der Linden, M., 2003. The contribution of recollection and familiarity to recognition memory: a study of the effects of test format and aging. Neuropsychology 17 (1), 14–24. https://doi.org/10.1037/0894-4105.17.1.14.

Bies-Hernandez, N.J., 2013. Examining the Testing Effect Using the Dual-Process Signal Detection Model. Doctoral Dissertation, University of Nevada, Las Vegas. Retrieved from. https://digitalscholarship.unlv.edu/thesesdissertations/1804.

Brown, P.C., Roediger, H.L., McDaniel, M.A., 2014. Make it Stick: the Science of Successful Learning. Harvard University Press, Cambridge, MA.

Buchin, Z.L., Mulligan, N.W., 2017. The testing effect under divided attention. J. Exp. Psychol. Learn. Mem. Cogn. 43 (12), 1934–1947. https://doi.org/10.1037/xlm0000427.

Carpenter, S.K., 2009. Cue strength as a moderator of the testing effect: the benefits of elaborative retrieval. J. Exp. Psychol. Learn. Mem. Cogn. 35 (6), 1563–1569. https://doi.org/10.1037/a0017021.

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

Chan, J.C.K., McDermott, K.B., 2007. The testing effect in recognition memory: a dual process account. J. Exp. Psychol. Learn. Mem. Cogn. 33 (2), 431–437. https://doi.org/10.1037/0278-7393.33.2.431.

Coane, J.H., 2013. Retrieval practice and elaborative encoding benefit memory in younger and older adults. J. Appl. Res. Memory Cogn. 2 (2), 95–100. https://doi.org/10.1016/j.jarmac.2013.04.001.

Conway, M.A., Gardiner, J.M., Perfect, T.J., Anderson, S.J., Cohen, G.M., 1997. Changes in memory awareness during learning: the acquisition of knowledge by psychology undergraduates. J. Exp. Psychol. Gen. 126 (4), 393–413. https://doi.org/10.1037/0096-3445.126.4.393.

Dewhurst, S.A., Conway, M.A., Brandt, K.R., 2009. Tracking the R-to-K shift: changes in memory awareness across repeated tests. Appl. Cognit. Psychol. 23 (6), 849–858. https://doi.org/10.1002/acp.1517.

Diana, R.A., Reder, L.M., Arndt, J., Park, H., 2006. Models of recognition: a review of arguments in favor of a dual-process account. Psychon. Bull. Rev. 13 (1), 1–21. https://doi.org/10.3758/BF03193807.

Diana, R.A., Yonelinas, A.P., Ranganath, C., 2013. Parahippocampal cortex activation during context reinstatement predicts item recollection. J. Exp. Psychol. Gen. 142 (4), 1287–1297. https://doi.org/10.1037/a0034029.

Dudukovic, N.M., DuBrow, S., Wagner, A.D., 2009. Attention during memory retrieval enhances future remembering. Mem. Cogn. 37 (7), 953–961. https://doi.org/10.3758/MC.37.7.953.

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. Psychol. Sci. Public Interest 14 (1), 4–58. https://doi.org/10.1177/1529100612453266.

Dunn, J.C., 2004. Remember–Know: a matter of confidence. Psychol. Rev. 111 (2), 524–542. https://doi.org/10.1037/0033-295X.111.2.524.

Eriksson, J., Kalpouzos, G., Nyberg, L., 2011. Rewiring the brain with repeated retrieval: a parametric fMRI study of the testing effect. Neurosci. Lett. 505 (1), 36–40. https://doi.org/10.1016/j.neulet.2011.08.061.

Gao, C., Rosburg, T., Hou, M., Li, B., Xiao, X., Guo, C., 2016. The role of retrieval mode and retrieval orientation in retrieval practice: insights from comparing recognition memory testing formats and restudying. Cognit. Affect Behav. Neurosci. 16 (6), 977–990. https://doi.org/10.3758/s13415-016-0446-z.

Gardiner, J.M., Java, R.I., 1991. Forgetting in recognition memory with and without recollective experience. Mem. Cogn. 19 (6), 617–623. https://doi.org/10.3758/BF03197157.

Gaspelin, N., Ruthruff, E., Pashler, H., 2013. Divided attention: an undesirable difficulty in memory retention. Mem. Cogn. 41 (7), 978–988. https://doi.org/10.3758/s13421-013-0326-5.

Hashimoto, T., Usui, N., Taira, M., Kojima, S., 2011. Neural enhancement and attenuation induced by repetitive recall. Neurobiol. Learn. Mem. 96 (2), 143–149. https://doi.org/10.1016/j.nlm.2011.03.008.

Jacoby, L.L., 1991. A process dissociation framework: separating automatic from intentional uses of memory. J. Mem. Lang. 30 (5), 513–541. https://doi.org/10.1016/0749-596X(91)90025-F.

Jang, Y., Wixted, J.T., Huber, D.E., 2009. Testing signal-detection models of yes/no and two-alternative forced-choice recognition memory. J. Exp. Psychol. Gen. 138 (2), 291–306. https://doi.org/10.1037/a0015525.

Jang, Y., Wixted, J.T., Huber, D.E., 2011. The diagnosticity of individual data for model selection: comparing signal-detection models of recognition memory. Psychon. Bull. Rev. 18 (4), 751–757. https://doi.org/10.3758/s13423-011-0096-7.

Jones, T.C., Roediger, H.L., 1995. The experiential basis of serial position effects. Eur. J. Cogn. Psychol. 7 (1), 65–80. https://doi.org/10.1080/09541449508520158.

Karpicke, J.D., Lehman, M., Aue, W.R., 2014. Retrieval-based learning: an episodic context account. In: Ross, B.H. (Ed.), Psychology of Learning and Motivation, vol. 61, pp. 237–284. https://doi.org/10.1016/B978-0-12-800283-4.00007-1.

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

Keresztes, A., Kaiser, D., Kovács, G., Racsmány, M., 2014. Testing promotes long-term learning via stabilizing activation patterns in a large network of brain areas. Cerebr. Cortex 24 (11), 3025–3035. https://doi.org/10.1093/cercor/bht158.

Kessler, Y., Vandermorris, S., Gopie, N., Daros, A., Winocur, G., Moscovitch, M., 2014. Divided attention improves delayed, but not immediate retrieval of a consolidated memory. PLoS One 9 (3), e91309. https://doi.org/10.1371/journal.pone.0091309.

Koen, J.D., Barrett, F.S., Harlow, I.M., Yonelinas, A.P., 2017. The ROC Toolbox: a toolbox for analyzing receiver-operating characteristics derived from confidence ratings. Behav. Res. Methods 49 (4), 1399–1406. https://doi.org/10.3758/s13428-016-0796-z.

Koen, J.D., Yonelinas, A.P., 2016. Recollection, not familiarity, decreases in healthy aging: converging evidence from four estimation methods. Memory 24 (1), 75–88. https://doi.org/10.1038/nmeth.2839.A.

Larsen, D.P., Butler, A.C., Roediger, H.L., 2008. Test-enhanced learning in medical education. Med. Educ. 42 (10), 959–966. https://doi.org/10.1111/j.1365-2923.2008.03124.x.

Lehman, M., Smith, M.A., Karpicke, J.D., 2014. Toward an episodic context account of retrieval-based learning: dissociating retrieval practice and elaboration. J. Exp. Psychol. Learn. Mem. Cogn. 40 (6), 1787–1794. https://doi.org/10.1037/xlm0000012.

Liu, X.L., Liang, P., Li, K., Reder, L.M., 2014. Uncovering the neural mechanisms underlying learning from tests. PLoS One 9 (3), e92025. https://doi.org/10.1371/journal.pone.0092025.

Liu, Y., Rosburg, T., Gao, C., Weber, C., Guo, C., 2017. Differentiation of subsequent memory effects between retrieval practice and elaborative study. Biol. Psychol. 127, 134–147. https://doi.org/10.1016/j.biopsycho.2017.05.010.

Logan, J.M., Balota, D.A., 2008. Expanded vs. equal interval spaced retrieval practice: exploring different schedules of spacing and retention interval in younger and older adults. Aging Neuropsychol. Cognit. 15 (3), 257–280. https://doi.org/10.1080/13825580701322171.

McCabe, D.P., Geraci, L.D., 2009. The influence of instructions and terminology on the accuracy of remember-know judgments. Conscious. Cognit. 18 (2), 401–413. https://doi.org/10.1016/j.concog.2009.02.010.

McDaniel, M.A., Thomas, R.C., Agarwal, P.K., McDermott, K.B., Roediger, H.L., 2013. Quizzing in middle-school science: successful transfer performance on classroom exams. Appl. Cognit. Psychol. 27 (3), 360–372. https://doi.org/10.1002/acp.2914.

McDermott, K.B., Agarwal, P.K., D'Antonio, L., Roediger, H.L., McDaniel, M.A., 2014. Both multiple-choice and short-answer quizzes enhance later exam performance in middle and high school classes. J. Exp. Psychol. Appl. 20 (1), 3–21. https://doi.org/10.1037/xap0000004.

Migo, E.M., Mayes, A.R., Montaldi, D., 2012. Measuring recollection and familiarity: improving the remember/know procedure. Conscious. Cognit. 21 (3), 1435–1455. https://doi.org/10.1016/j.concog.2012.04.014.

Mulligan, N.W., Picklesimer, M., 2016. Attention and the testing effect. J. Exp. Psychol. Learn. Mem. Cogn. 42 (6), 938–950. https://doi.org/10.1037/xlm0000227.

Parks, C.M., Yonelinas, A.P., 2007a. Moving beyond pure signal-detection models: comment on Wixted (2007). Psychol. Rev. 114 (1), 188–202. https://doi.org/10.1037/0033-295X.114.1.188.

Parks, C.M., Yonelinas, A.P., 2007b. Postscript: comment on wixted (2007). Psychol. Rev. 114 (1), 201–202. https://doi.org/10.1037/0033-295X.114.1.201.

Peng, Y., Liu, Y., Guo, C., 2019. Examining the neural mechanism behind testing effect with concrete and abstract words. Neuroreport 30 (2), 113–119. https://doi.org/10.1097/WNR.0000000000001169.

Pu, X., Tse, C.-S., 2014. The influence of intentional versus incidental retrieval practices on the role of recollection in test-enhanced learning. Cogn. Process. 15 (1), 55–64. https://doi.org/10.1007/s10339-013-0580-2.

Rabinowitz, J.C., Craik, F.I.M., 1986. Prior retrieval effects in young and old adults. J. Gerontol. 41 (3), 368–375. https://doi.org/10.1093/geronj/41.3.368.

Ranganath, C., 2010. Binding items and contexts: the cognitive neuroscience of episodic memory. Curr. Dir. Psychol. Sci. 19 (3), 131–137. https://doi.org/10.1177/0963721410368805.

Roediger, H.L., Agarwal, P.K., McDaniel, M.A., McDermott, K.B., 2011. Test-enhanced learning in the classroom: long-term improvements from quizzing. J. Exp. Psychol. Appl. 17 (4), 382–395. https://doi.org/10.1037/a0026252.

Roediger, H.L., Butler, A.C., 2011. The critical role of retrieval practice in long-term retention. Trends Cogn. Sci. 15 (1), 20–27. https://doi.org/10.1016/j.tics.2010.09.003.

Roediger, H.L., Karpicke, J.D., 2006a. Test-enhanced learning: taking memory tests improves long-term retention. Psychol. Sci. 17 (3), 249–255. https://doi.org/10.1111/j.1467-9280.2006.01693.x.

Roediger, H.L., Karpicke, J.D., 2006b. The power of testing memory: basic research and implications for educational practice. Perspect. Psychol. Sci. 1 (3), 181–210. https://doi.org/10.1111/j.1745-6916.2006.00012.x.

Rosburg, T., Johansson, M., Weigl, M., Mecklinger, A., 2015. How does testing affect retrieval-related processes? An event-related potential (ERP) study on the short-term effects of repeated retrieval. Cognit. Affect Behav. Neurosci. 15 (1), 195–210. https://doi.org/10.3758/s13415-014-0310-y.

Rotello, C.M., Macmillan, N.A., Hicks, J.L., Hautus, M.J., 2006. Interpreting the effects of response bias on remember–know judgments using signal detection and threshold models. Mem. Cogn. 34 (8), 1598–1614. https://doi.org/10.3758/BF03195923.

Rotello, C.M., Macmillan, N.A., Reeder, J.A., Wong, M., 2005. The remember response: subject to bias, graded, and not a process-pure indicator of recollection. Psychon. Bull. Rev. 12 (5), 865–873. https://doi.org/10.3758/BF03196778.

Rowland, C.A., 2011. Testing Effects in Context Memory. Master's Thesis. Colorado State University. Retrieved from. https://mountainscholar.org/bitstream/handle/10217/46908/Rowland_colostate_0053N_10630.pdf?sequence=1&isAllowed=y.

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

Skinner, E.I., Fernandes, M.A., 2007. Neural correlates of recollection and familiarity: a review of neuroimaging and patient data. Neuropsychologia 45 (10), 2163–2179. https://doi.org/10.1016/j.neuropsychologia.2007.03.007.

Tse, C.-S., Balota, D.A., Roediger, H.L., 2010. The benefits and costs of repeated testing on the learning of face-name pairs in healthy older adults. Psychol. Aging 25 (4), 833–845. https://doi.org/10.1037/a0019933.

Tulving, E., 1967. The effects of presentation and recall of material in free-recall learning. J. Verb. Learn. Verb. Behav. 6 (2), 175–184. https://doi.org/10.1016/S0022-5371(67)80092-6.

Tulving, E., 1985. Memory and consciousness. Canadian Psychol./ Psychol. Canadienne 26 (1), 1–12. https://doi.org/10.1037/h0080017.

van den Broek, G., Takashima, A., Segers, E., Fernández, G., Verhoeven, L., 2013. Neural correlates of testing effects in vocabulary learning. Neuroimage 78, 94–102. https://doi.org/10.1016/j.neuroimage.2013.03.071.

van den Broek, G., Takashima, A., Wiklund-Hörnqvist, C., Karlsson Wirebring, L., Segers, E., Verhoeven, L., Nyberg, L., 2016. Neurocognitive mechanisms of the "testing effect": a review. Trends Neurosci. Edu. 5 (2), 52–66. https://doi.org/10.1016/j.tine.2016.05.001.

Verkoeijen, P.P.J.L., Tabbers, H.K., Verhage, M.L., 2011. Comparing the effects of testing and restudying on recollection in recognition memory. Exp. Psychol. 58 (6), 490–498. https://doi.org/10.1027/1618-3169/a000117.

Vilberg, K.L., Rugg, M.D., 2007. Dissociation of the neural correlates of recognition memory according to familiarity, recollection, and amount of recollected information. Neuropsychologia 45 (10), 2216–2225. https://doi.org/10.1016/j.neuropsychologia.2007.02.027.

Wing, E.A., Marsh, E.J., Cabeza, R., 2013. Neural correlates of retrieval-based memory enhancement: an fMRI study of the testing effect. Neuropsychologia 51 (12), 2360–2370. https://doi.org/10.1016/j.neuropsychologia.2013.04.004.

Wixted, J.T., 2007a. Dual-process theory and signal-detection theory of recognition memory. Psychol. Rev. 114 (1), 152–176. https://doi.org/10.1037/0033-295X.114.1.152.

Wixted, J.T., 2007b. Spotlighting the probative findings: reply to Parks and Yonelinas (2007). Psychol. Rev. 114 (1), 203–209. https://doi.org/10.1037/0033-295X.114.1.203.

Wixted, J.T., Mickes, L., 2010. A continuous dual-process model of remember/know judgments. Psychol. Rev. 117 (4), 1025–1054. https://doi.org/10.1037/a0020874.

Yonelinas, A.P., 1994. Receiver-operating characteristics in recognition memory: evidence for a dual-process model. J. Exp. Psychol. Learn. Mem. Cogn. 20 (6), 1341–1354. https://doi.org/10.1037/0278-7393.20.6.1341.

Yonelinas, A.P., 2001. Consciousness, control, and confidence: the 3 Cs of recognition memory. J. Exp. Psychol. Gen. 130 (3), 361–379. https://doi.org/10.1037/0096-3445.130.3.361.

Yonelinas, A.P., 2002. The nature of recollection and familiarity: a review of 30 years of research. J. Mem. Lang. 46 (3), 441–517. https://doi.org/10.1006/JMLA.2002.2864.

Yonelinas, A.P., Aly, M., Wang, W.-C., Koen, J.D., 2010. Recollection and familiarity: examining controversial assumptions and new directions. Hippocampus 20 (11), 1178–1194. https://doi.org/10.1002/hipo.20864.

Yonelinas, A.P., Jacoby, L.L., 1995. The relation between remembering and knowing as bases for recognition: effects of size congruency. J. Mem. Lang. 34 (5), 622–643. https://doi.org/10.1006/jmla.1995.1028.

Yonelinas, A.P., Levy, B.J., 2002. Dissociating familiarity from recollection in human recognition memory: different rates of forgetting over short retention intervals. Psychon. Bull. Rev. 9 (3), 575–582. https://doi.org/10.3758/BF03196315.

Yonelinas, A.P., Otten, L.J., Shaw, K.N., Rugg, M.D., 2005. Separating the brain regions involved in recollection and familiarity in recognition memory. J. Neurosci. 25 (11), 3002–3008. https://doi.org/10.1523/JNEUROSCI.5295-04.2005.

Yonelinas, A.P., Parks, C.M., 2007. Receiver operating characteristics (ROCs) in recognition memory: a review. Psychol. Bull. 133 (5), 800–832. https://doi.org/10.1037/0033-2909.133.5.800.