Memory & Cognition
https://doi.org/10.3758/s13421-019-00997-3

Long-term inference and memory following retrieval practice

Jessica Siler[1] · Aaron S. Benjamin[1,2]

Abstract
One exceptional characteristic of the testing effect is its generalizability over time and circumstance. The benefits of testing over rote restudy appear to grow with time, as forgetting occurs, and also have been documented to extend to tasks of inference on previously unstudied stimuli. In the two experiments reported here, we evaluated inference and memory for members of natural categories over time. Rote memory and generalization were tested shortly after the study phase and again after varying delays. Results from both experiments indicate that retrieval practice does indeed enhance inference for novel members of previously learned categories, and that the benefits are maintained over the duration of our experiments—up to 25 days. An analysis of forgetting rates indicates that retrieval practice does not, however, decelerate forgetting when compared with restudy. Rates of forgetting were not discernibly different, for either rote memory or conceptual knowledge, between the two conditions. These results indicate that although testing does not appear to reduce forgetting, it is a potent means of enhancing inference, and the benefits to memory and inference are long lasting.

Keywords Testing effect · Transfer · Forgetting

The benefits to memory of the testing effect are now well known within cognitive psychology. Retrieving information enhances memory for that information more than passive restudy (for review, see Rowland, 2014). There is evidence that the benefits of testing extend beyond rote memory, and beyond the short time scales used in traditional laboratory research. Testing appears to reduce the rate with which we forget information (Carpenter, Pashler, Wixted, & Vul, 2008; Roediger & Karpicke, 2006a; Runquist, 1983; Wheeler, Ewers, & Buonanno, 2003; Wheeler & Roediger, 1992) and to improve inference and transfer on materials that go beyond the tested stimuli (Butler, 2010; Jacoby, Wahlheim, & Coane, 2010). Taken together, the versatility of the testing effect suggests high potential value as an educational tool (Roediger & Karpicke, 2006b; Benjamin & Pashler, 2015; Karpicke, 2012; Roediger, Agarwal, McDaniel, & McDermott, 2011a; Roediger, Putnam, & Smith, 2011b).

✉ Jessica Siler
siler3@illinois.edu

[1] Department of Psychology, University of Illinois at Urbana-Champaign, 603 E. Daniel St, Champaign, IL 61820, USA

[2] Beckman Institute for Advanced Science and Technology, University of Illinois at Urbana-Champaign, Champaign, IL, USA

These many sources of evidence notwithstanding, testing is not widely used in educational settings (Benjamin & Pashler, 2015). In many college courses, tests are used exclusively as assessment tools. One challenge faced by promoters of testing as a learning tool is the poor intuition students and teachers have about nature of tests and of human memory (Bjork, Dunlosky, & Kornell, 2013). When surveyed about their study habits, a majority of university students reported studying only by rereading material, not by testing themselves. Students fail to predict the many benefits of testing (Karpicke, Butler, & Roediger, 2009; Roediger & Karpicke, 2006a; Roediger, Putnam, et al., 2011b; Tullis, Fiechter, & Benjamin, 2018; Tullis, Finley, & Benjamin, 2013), and parents and teachers decry the widespread growing role of standardized tests in the classroom ("The 47th PDK/Gallup Poll," 2015).

Students' preference for restudying over testing is likely due to a combination of factors. They may experience an enhancement of perceptual fluency with repeated rereading that fosters a sense of confidence, but does not predict retrievability (Benjamin, Bjork, & Schwartz, 1998; Jacoby & Dallas, 1981; Morris, Bransford, & Franks, 1977). More generally, there seems to be a widespread lack of understanding of the nature of tests and about the malleability of human memory (Loftus, 2003), leading learners to underappreciate the fact that tests have the potential to change, rather than merely assess, memory. There are a variety of misleading metaphors

Published online: 09 December 2019

in popular circulation: memory as a storehouse (Goldsmith & Koriat, 2007); tests as mirrors that reflect what we know (Benjamin & Pashler, 2015); and encoding as the key to the construction of knowledge (Karpicke, 2012). In reality, memory is quite fluid and can be shaped through the retrieval practice offered by tests.

In the following two sections, we critically review what is known about two questions that are central to a theoretical understanding of the testing effect and to the application of the effect in the classroom and beyond. Those two questions are: *Does testing slow the rate of forgetting?* and *Does testing enhance inference and generalization to new materials*? We then report two experiments that evaluate memory and transfer over extended delays. These experiments allow us to examine forgetting more thoroughly than has been done in previous research, including the critical question of whether enhancements to inference persevere over long delays.

Testing and forgetting

Some reports on the testing effect have concluded that testing benefits memory by diminishing the rate of forgetting (Runquist, 1983; Thompson, Wenger, & Bartling, 1978). This conclusion is compatible with the finding that, whereas the benefits of testing after 1 day are robust, they are often absent or even reversed after short intervals. In some studies, the testing effect appears to grow even more with yet longer intervals (Carpenter et al., 2008; Roediger & Karpicke, 2006a; Toppino & Cohen, 2009; Wheeler et al., 2003).

However, drawing inferences about forgetting rates from examining performance over time is a subtle endeavor. There are substantial individual differences in forgetting rates (Kyllonen & Tirre, 1988), and it is well known that averaging over individuals with different rates can yield a function best fit with a parametric form different from the individual contributors (Anderson, 2001; Estes, 1956; Murre & Chessa, 2011). This means that forgetting functions averaged over individuals can be qualitatively and quantitatively misleading. In addition, test formats that allow participants to remember the same item more than once across different tests introduce the possibility that retrieval on an earlier test will affect memory on a later test (e.g., Roediger, Putnam, et al., 2011). The beneficial confounding effects of earlier tests on later ones has the potential to mimic the effects of slowed forgetting. Together, these factors constrain the types of experiments that can be used to meaningfully measure forgetting rates following testing.

Wheeler et al. (2003) reported two experiments in which participants studied lists of nouns, either by repeated restudy or by a single study event followed by repeated tests. They observed the traditional interaction: participants in the restudy condition performed best on an immediate free recall test (after 5 min), but the participants in the testing condition performed better after 1 week. They concluded that forgetting is faster following restudy than following a test. However, because each test in their procedure invited recall of all the memoranda via a free recall test, there is little doubt that the differential benefits to memory from testing and restudy on the first test affected memory on the second test. Thus, although it is a valid conclusion that repeated testing can yield relatively greater retention with time, it is not evident from the study of Wheeler et al. that it does so by retarding forgetting. It may simply be that, as discussed above, each test adds more strength to memory, yielding slower *net* forgetting.

A later study provides a stronger assessment of whether testing retards forgetting, and strongly influenced the design we present here. In a study by Carpenter et al. (2008), participants studied obscure facts or Swahili-English word pairs and were tested immediately (after 5 min) and after 1, 2, 7, 14, and 42 days. Because they measured memory at multiple time points, within subjects, they were able to fit each individual's data to a single forgetting function and estimate the rate of forgetting from that function. In addition, they used a cued-recall procedure that controlled which items were tested at each interval, ensuring that no item was tested more than once. Such a technique yields a truer forgetting curve for an individual. They did find that testing retarded forgetting, but the effects were small in magnitude and not nearly as dramatic as those evident in Wheeler et al. Our experiments provide an opportunity to corroborate this result in a similar design and to extend the question about reduced forgetting beyond rote retention to conceptual knowledge.

Testing and transfer

Only recently have we appreciated that testing can yield benefits beyond those evident on traditional tests of rote memory and can support more meaningful learning as well (Pan & Rickard, 2018). Compared with continued restudy, testing can foster better understanding of psychological concepts (Wiklund-Hörnqvist, Jonsson, & Nyberg, 2014), lead to higher scores on statistics exams with novel problems (Lyle & Crawford, 2011), and improve skill learning (Kromann, Jensen, & Ringsted, 2009). This extension of the testing effect is important—the goal of learning is not only to promote high levels of performance in the long term, but also to promote transfer to related tasks and contexts (Schmidt & Bjork, 1992).

A limited number of studies have addressed the effects of testing on transfer across knowledge domains. These studies show that the memorial benefits of testing stretch beyond the specific responses from the initial tests. Butler (2010) showed that tests on text passages can lead to superior transfer on questions of inference. Similarly, Chan (2010) showed that tests can facilitate recall of studied material that was never

tested. Testing has also yielded transfer in the context of map learning (Rohrer, Taylor, & Sholar, 2010) and function rule learning (Kang, McDaniel, & Pashler, 2011). Jacoby et al. (2010) provided a particularly good example with concept learning. In three experiments, participants studied pictures of birds and the families to which they belonged. In the restudy condition, the bird picture was continuously presented with the family name. In the retrieval practice condition, the picture of the bird was presented, and the participant was prompted to remember the correct family name. After their guess, they were provided with the correct answer. Either immediately after this study phase or on the next day, participants completed the final test in which they were tested on the set of birds they encountered during study and a completely new set of birds from the same families. The retrieval practice condition led to superior categorization of both studied *and* novel birds on both tests.

The experiments we report here closely matched the procedure from Experiment 3 from Jacoby et al. (2010). One goal of this work was to replicate and extend Jacoby et al.'s very important finding that retrieval practice leads to superior inference on new materials. A second goal of these experiments, featured in Experiment 2, was to precisely measure forgetting on a recognition test (for studied materials) and inference (for new materials), and to provide a new test of whether testing retards forgetting. In combining an assessment of forgetting with tests of inference, we can also evaluate the novel question of whether performance on tests of inference reveal slower forgetting of conceptual knowledge following retrieval practice.

Experiment 1

The purpose of Experiment 1 was to replicate and extend the findings of Jacoby et al. (2010) using a similar category learning task. This experiment closely matches Experiment 3 of Jacoby et al. but uses a 1-week delay, has fewer study opportunities, and is completely within subjects. Study materials and data can be found on our OSF project page (https://osf.io/suqcx/).

Method

Participants Seventy-five undergraduate students from the University of Illinois at Urbana-Champaign participated in exchange for partial course credit. Data from nine additional participants were eliminated from analysis due to technical problems or failure to complete the experiment. Performance from the eliminated participants was not measurably different from that of the 75 included in the analyses. Mean performance across the three rounds of the study phase for the eliminated participants ($M_{R1} = 0.43$, $M_{R2} = 0.65$, $M_{R3} = 0.79$) was higher than for those participants kept in the following analyses ($M_{R1} = 0.43$, $M_{R2} = 0.57$, $M_{R3} = 0.69$). Therefore, it is unlikely that these participants dropped out of the study due to poor performance (cf. Steyvers & Benjamin, 2018).

Materials A total of 128 bird images were used as stimuli in this experiment. Sixteen images were selected for each of the following eight bird families: finch, jay, warbler, chickadee, flycatcher, bunting, thrush, and swallow. These bird families all originate from the same taxonomic order (*Passeriformes*). Following Jacoby et al. (2010), families from this order were chosen to ensure enough between-family similarity to make the categorization task sufficiently difficult. The images were gathered from www.whatbird.com. All 128 color images were presented at the same size and featured each bird in a perching position.

Design This experiment employed a 2 × 2 × 2 within-subjects design. We manipulated study condition (restudy or retrieval practice), exemplar (studied or novel), and retention interval (immediate or 7 days). Of the eight bird families learned, four were assigned to be repeatedly restudied and four to be retrieved. In the restudy condition, each bird image was presented at the center of the screen with the family name printed underneath. In the retrieval practice condition, each bird image was presented at the center of the screen alone, and participants were required to enter the family name in a blank text box underneath. Six images from each of the eight bird families were randomly assigned to be studied, and the remaining 10 from each family were assigned to be novel. The 48 studied images were seen during the study phase and final tests of the experiment, and the 80 novel images were seen only on the final tests. Participants completed two tests after the study phase of the experiment. The first test was completed immediately after the study phase; the second test was completed 7 days later. Each test included images from all eight families, but no images were repeated between tests. A total of 64 images were used per test: three studied and five novel birds from each of the eight families. The families and exemplars assigned to each condition were randomized for each participant.

Procedure Participants were told they would be studying a variety of birds and the families to which they belong, and that they would be tested on what they have learned. The full experiment was conducted in the lab and consisted of a training phase, study phase, and two tests.

Participants began with a brief training phase, the purpose of which was to familiarize them with the birds and the families they would be learning. In training, all 48 unique bird images from the studied set were presented in random order. Each image was presented for 4 s at the center of the screen, accompanied by its family name.

Participants then advanced to the study phase and were told they would have three more opportunities to study those birds and their families. From this point, half of the bird families were restudied, and half were retrieved. On restudy trials, each bird image was presented along with its family name, and participants were instructed to retype the name on each trial. Forcing learners to type the name even in the restudy condition ensures that the differences in conditions are not due to differences in preparing and executing a response (as in the production effect; MacLeod, Gopie, Hourihan, Neary, & Ozubko, 2010). Participants were given 4 s to type in the name; the image and name remained on the screen for an additional 1 s after they submitted their response. On retrieval practice trials, each bird image was presented alone, and participants were instructed to evaluate to which family that bird belonged, and they would be told if they were right or wrong and given the correct family name. Participants were given approximately 5 s to type in their guess and were then provided with corrective feedback that remained on the screen for an additional 2 s. If participants took longer than the 4–5 s to enter their responses, a message appeared that instructed them to try to answer more quickly on future trials. This time restriction was implemented in an effort to roughly equate time on task in the two conditions. Though retrieval practice trials were longer, participants experienced only 2 s with the correct information presented, compared with 5 s on the restudy trials. After completing the first round of practice with all 48 birds, participants completed two more rounds under the same conditions. The order in which the 48 trials were presented in each round was randomized.

After the study phase, participants proceeded to the first of two tests. Participants were told that the test would include the birds they just studied along with new birds from the same families. The test evaluated both categorization and recognition of the birds. Each test item involved a bird image, and participants first had to evaluate to which family it belonged (categorization), then had to decide whether they had seen that exact image previously in the experiment (recognition). Throughout the test, participants were provided with all eight family names at the bottom on the screen. We were concerned that the family names would not be as available in memory as they were during the study phase, especially after long delays, so we provided the names on the screen to avoid any omissions or typos on the final tests. After completing the first test, participants were dismissed and returned 7 days later to complete the second test, which was of the same format as the first. No feedback was provided on either test.

Results

Study phase Performance across the three rounds of study for the retrieval practice birds was analyzed via one-way ANOVA. Study phase performance significantly improved over the three rounds, $F(2, 148) = 183.3$, $p < .001$ (M_{R1}= 0.43, $M_{R2} = 0.57$, $M_{R3} = 0.69$). This result demonstrates that participants learned the identities of the studied birds over the course of the experiment.

Recognition Discriminability (d') was calculated from each participant's data and compared in a 2 × 2 (study condition × retention interval) repeated-measures ANOVA. ANOVA is an acceptable tool for analyzing the data here because, unlike data expressed as proportions or response times, d' has an interpretation that affords treatment on a linear scale. Performance using d' is shown in Fig. 1. There were significant main effects of both study condition, $F(1, 74) = 63.37$, $p < .001$, and retention interval, $F(1, 74) = 93.72$, $p < .001$. These results indicate that birds from retrieved families were better remembered than birds from restudied families, and that memory was superior on the immediate than the delayed test. The interaction was not statistically significant. The absolute drop in forgetting between the two tests was greater for retrieved families (0.76) than restudied families (0.57), a result that is inconsistent with the claim that forgetting is slower following retrieval practice. However, this experiment was not designed to measure forgetting rates and does not provide the multiple time points necessary to do so accurately. Experiment 2 provides a better opportunity to make this comparison.

Categorization accuracy Categorization performance on the final tests was compared in two 2 × 2 (study condition × retention interval) repeated-measures ANOVAs for the studied and novel items. Main effects were found for study condition in both the studied and novel sets of items, $F(1, 74) = 59.27$, $p < .001$ and $F(1, 74) = 12.59$, $p < .001$, respectively. Overall, this indicates that retrieval practice led to better categorization accuracy than restudy at both the long and short retention interval. In addition, a main effect of retention

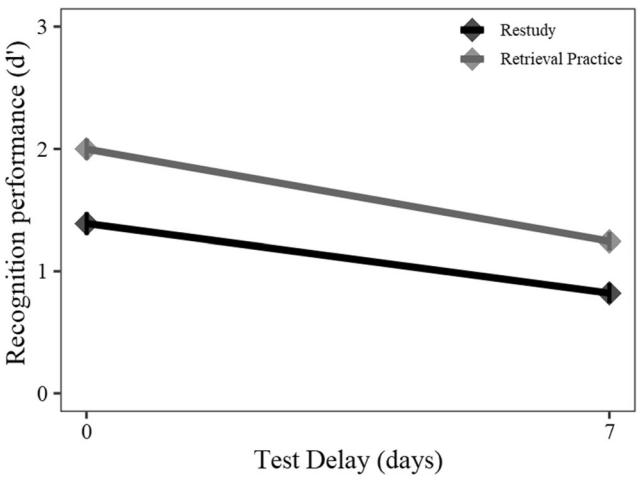

Fig. 1 Recognition of birds over test delay and study condition (error bars represent the standard error; Experiment 1)

interval was found in the studied set of items, $F(1, 74) = 50.79$, $p < .001$, but not in the novel items. Looking at the data in Fig. 2, it appears that categorization of the studied items was affected by retention interval, but the novel items were not. This result is consistent with the idea that previously viewed exemplars were sometimes classified by memory for the specific stimulus, rather than knowledge of the category, and that it is that source of evidence that underwent measurable forgetting over the delay.

Discussion

The results from Experiment 1 replicated the critical findings of Jacoby et al. (2010): retrieval practice enhanced memory for studied exemplars and led to superior categorization of stimuli. Most importantly, the categorization advantage extended to totally novel stimuli from studied categories, and that advantage lasted at least 1 week.

Experiment 1 revealed no differences in forgetting rates between restudied and retrieved items over the course of 1 week. However, the experiment is poorly designed for such measurement: It includes only two time points and does not extend the retention interval to a distant point with ample forgetting. In order to assess whether differential forgetting occurs between conditions, an experimental design with multiple time points is required. An experimental design with just two points in time may hint at differences, but cannot definitively demonstrate these due to an underlying linear assumption built into statistical tests like ANOVAs. An interaction between condition and retention interval may falsely suggest differential forgetting; likewise, a null finding using a linear model may disguise different forgetting rates. Since forgetting is well categorized by a power-law function (e.g., Rubin & Wenzel, 1996; Wickens, 1998; Wixted, 2004), a design with multiple time points is needed to fit forgetting functions to individual participant's data and directly compare forgetting rates across study conditions, much like Carpenter et al.'s (2008). The extended intervals also provide an opportunity to assess the benefits to inference from retrieval practice over an even longer period, and to assess forgetting rates for categorization.

Experiment 2

The purpose of Experiment 2 was to replicate the findings of Experiment 1 and examine forgetting more closely. Experiment 2 included four final tests that extended out nearly a month. This allowed us to examine the effects of study condition on memory and inference at extremely long delays. Also, it allowed us to more directly investigate the effect of study condition on forgetting rates by collecting enough data to fit forgetting functions for each individual participant.

Method

Participants Seventy undergraduate students from the University of Illinois at Urbana-Champaign participated in exchange for partial course credit. Data from 19 additional participants were eliminated from analysis because of failure to complete the full experiment. The eliminated participants' performance did not differ substantially from that of the included participants. Though mean performance across both rounds of the study phase for the eliminated participants ($M_{R1} = 0.51$, $M_{R2} = 0.65$) was lower than for those participants kept in ($M_{R1} = 0.57$, $M_{R2} = 0.71$), they experienced roughly the same degree of learning across rounds. Therefore, though there is some cause for concern over their overall low performance, it is unlikely these participants

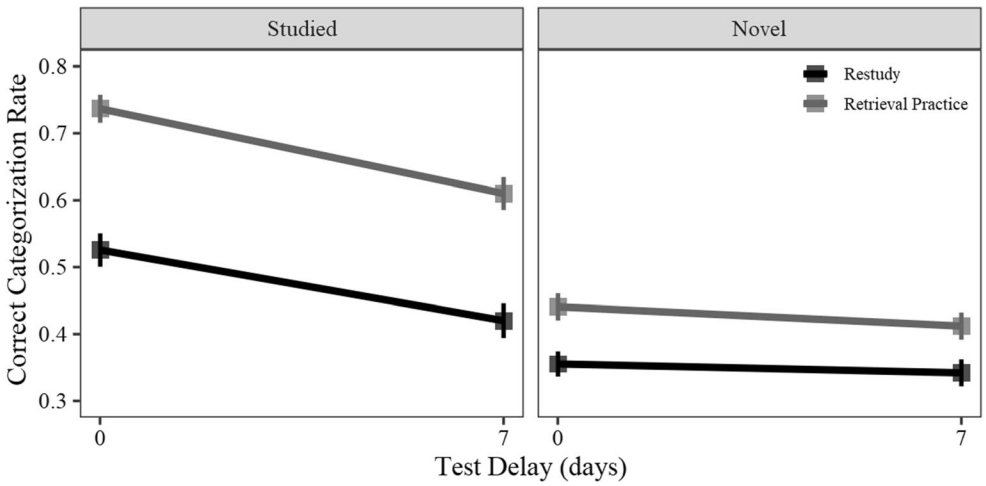

Fig. 2 Average categorization of studied and novel sets of birds across test delay and study conditions (error bars represent the standard error; Experiment 1)

dropped out of the study because of little to no learning in the study phase.

Materials A total of 200 bird images were used as stimuli in this experiment. Twenty images were selected for each of the following 10 bird families: finch, jay, chickadee, flycatcher, bunting, thrush, swallow, wren, oriole, and starling. As in Experiment 1, these bird families originate from the same taxonomic order (*Passeriformes*). The 200 images were presented in the same manner as in Experiment 1.

Design This experiment employed a 2 × 2 × 4 within-subjects design. We manipulated study condition (restudy or retrieval practice), exemplar (studied or novel), and retention interval (immediate, after 1 day, after 7 days, or after 25 days). Of the 10 bird families learned, five were assigned to be repeatedly restudied, and five to retrieval practice. Eight images from each of the 10 bird families were randomly assigned to be studied, and the remaining 12 from each family were assigned to be novel. The 80 studied images were seen during the study phase and final tests, and the 120 novel images were only seen on the final tests. Participants completed four tests after the study phase of the experiment. The first test was completed immediately after the study phase; the second upon return 1 day later; the third 7 days later; and the fourth approximately 25 days later. Each test included images from all eight families, but no images were repeated between tests. A total of 50 images were used per test: two studied and three novel from each of the 10 families. The families and exemplars assigned to each condition were randomized for each participant.

Procedure The procedure for Experiment 2 closely matched that of Experiment 1. The main differences were the size of the item sets, the duration of the study phase, and the number of tests. Given that we expanded the number of categories participants will learn in Experiment 2, we reduced the length of the study phase to just two rounds of practice so that the experiment would not run over the allotted time.

Results

Study phase Performance across the two rounds of study for the retrieval practice birds was analyzed via paired *t*-test. Study phase performance significantly improved from Round 1 ($M = 0.57$) to Round 2 ($M = 0.71$), $t(69) = 13.93$, $p < .001$. This result demonstrates that participants learned the training stimuli over the course of the experiment.

Recognition Discriminability (d') was calculated from each participant's data. A 2 × 4 (study condition × retention interval) repeated-measures ANOVA was used to analyze the results inferentially. As in Experiment 1, there were significant main effects of both study condition, $F(1, 69) = 55.26$, $p < .001$, and retention interval, $F(1, 69) = 35.49$, $p < .001$. Categories learned via retrieval practice were remembered better than ones repeatedly restudied; in addition, forgetting occurred over the course of the experiment. These effects are shown in Fig. 3. The interaction was not statistically significant.

Categorization accuracy Categorization performance on the final tests was compared in two 2 × 4 (study condition × retention interval) repeated-measures ANOVAs for the studied and novel items. Main effects were found for study condition in both the studied and novel sets of items, $F(1, 69) = 37.24$, $p < .001$, and $F(1, 69) = 15.24$, $p < .001$, respectively. Once again, retrieval practice led to better categorization accuracy then restudy. Within the studied set of items, there was a main effect of retention interval, $F(1, 69) = 11.79$, $p < .001$, indicating there was forgetting over the course of the experiment. Additionally, there was an interaction between study condition and retention interval, $F(1, 69) = 3.29$, $p = .022$, for the studied items. Looking at the data in Fig. 4, it appears that categorization accuracy of the studied items that were retrieved were most affected by the delay of the test. This finding also replicates results that were apparent in Experiment 1. No other effects were statistically significant.

Forgetting To assess condition effects on the rate of forgetting, we fit each individual participant's recognition performance and categorization performance to a power-law forgetting function: $y = \mathbf{a}(\mathbf{b}t + 1)^{-\mathbf{c}}$ (first described in Wickelgren, 1974). In this function, a represents the original degree of learning, b is a scaling parameter, c represents the forgetting rate, and t is time (in days). Individual forgetting functions were fit to the recognition and categorization data using maximum likelihood estimation. Additionally, average forgetting functions for each study condition were calculated based on median parameter estimates.

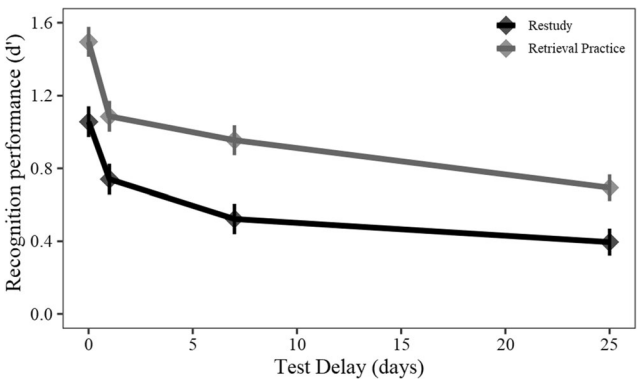

Fig. 3 Recognition of birds over test delay and study condition (error bars represent the standard error; Experiment 2)

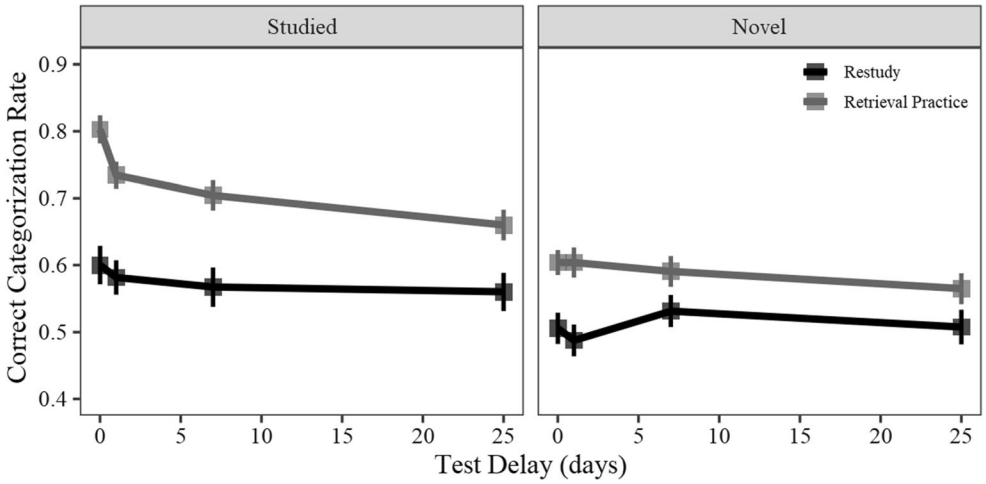

Fig. 4 Average categorization of studied and novel sets of birds across test delay and study conditions (error bars represent the standard error; Experiment 2)

Recognition Before fitting the individual forgetting functions, a general forgetting function was fit to the averages of all the recognition data. The *b* parameter estimated from this general function was used as a scaling constant for all the individual functions. The *a* and *c* parameters were then estimated for each participant across both conditions. To prevent extreme parameter estimates, the original degree of learning (*a*) parameter was restricted to a *d′* range of 0 to 4, and the forgetting rate parameter (*c*) was restricted to a range of 0 to 1. Most participants' data was well fit even without the imposition of parameter restrictions, but some cases produced uninterpretable parameter estimates.

The median values of *a* and *c* for the two learning conditions are shown in Table 1. To account for the paired structure in the data and for the likely nonnormal distributions, conditional effects on the parameters were compared using a binomial sign test. Categories that were retrieved revealed a higher original degree of learning than categories that were restudied for most participants (64%, $p = .022$). However, as shown in Fig. 5, there was no significant difference in forgetting rates between the restudy and retrieval practice conditions. Given that this result did not support the original hypothesis, we assert that the evidence here does not support the claim that retrieval practice slows forgetting. To evaluate the strength of evidence in favor of the null hypothesis, we conducted a Bayesian version of the binomial sign test (Morey & Rouder, 2015). These tests echoed the finding that retrieved categories led to higher original degree of learning over restudied categories ($BF_{10} = 4.02$). Yet, it remains unclear whether the two conditions differed in their effects on forgetting rates but leans towards the null hypothesis of no difference ($BF_{10} = 0.43$).

Table 1 Median estimates of *a* and *c* for each study condition for the recognition data (scaling parameter, $b = 8.88$; Experiment 2)

	Original degree of learning (*a*)	Forgetting rate (*c*)
Restudy	1.01	0.24
Retrieval practice	2.38	0.19

Categorization accuracy The same fitting routine was applied to data from the categorization task, separately for studied and novel items. The original degree of learning (*a*) parameter was restricted to a range of 0 to 1, and the forgetting rate parameter (*c*) was restricted to a range of 0 to 1. The median values of *a* and *c* for the two learning conditions are shown in Table 2 and plotted in Fig. 6. Because previously studied items can be categorized either on the basis of category knowledge or by rote memory from the studied exemplar, the novel items provide the purest measure of forgetting rates for purely categorical knowledge.

For previously studied items, most participants showed higher original degrees of learning for material that was retrieved than material that was restudied (80%, $p < .001$), as shown in Fig. 7. There was no significant difference in forgetting rates between the conditions, $p = .07$, and the Bayesian binomial sign test did not reveal interpretable effects ($BF_{10} = 1.54$).

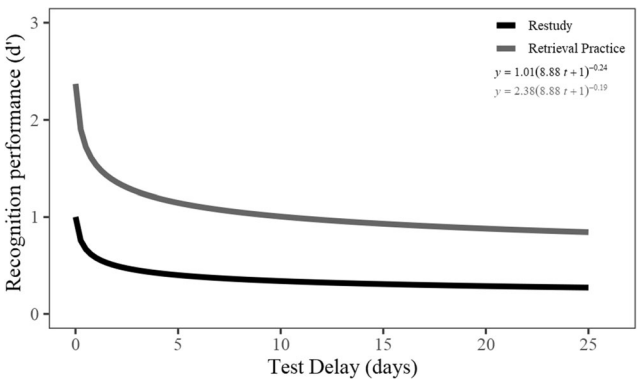

Fig. 5 Forgetting functions fit to recognition data based on median parameter estimates for each study condition (Experiment 2)

Table 2 Median estimates of a and c for each study condition and exemplar type for categorization (scaling parameter, b = 9.71; Experiment 2)

		Original degree of learning (a)	Forgetting rate (c)
Studied	Restudy	0.65	0.007
	Retrieval practice	0.85	0.037
Novel	Restudy	0.53	≈ 0
	Retrieval practice	0.64	0.020

The novel exemplars from categories that were retrieved elicited higher initial performance than ones from categories that were restudied (67%, p = .006). However, Fig. 8 shows that there was no advantage for either condition when comparing the rates of forgetting for categorization of novel stimuli. The results of the Bayesian binomial sign test support the conclusion that retrieval practice led to higher original degrees of learning than restudy (BF_{10} = 13.19), but differences in forgetting rates between retrieval practice and restudy were ambiguous (BF_{10} = 1.03).

Discussion

Results from Experiment 2 were entirely consistent with those of Experiment 1. Retrieval practice produced higher discriminability and higher categorization performance across all four tests. Although the advantage of retrieval practice over restudy was most pronounced for the set of studied items, the tendency generalized to novel items, consistent with the findings of Jacoby et al. (2010).

However, the analysis of forgetting rates did not replicate previous research. It has been claimed that retrieval practice slows forgetting, but there was no evidence for such an effect in either experiment, for either recognition or categorization. From these analyses, it appears that retrieval practice leads to higher initial degree of learning, but no benefits in decreased forgetting. Evidence in support of the *absence* of such effects was also weak, however, rendering this conclusion provisional.

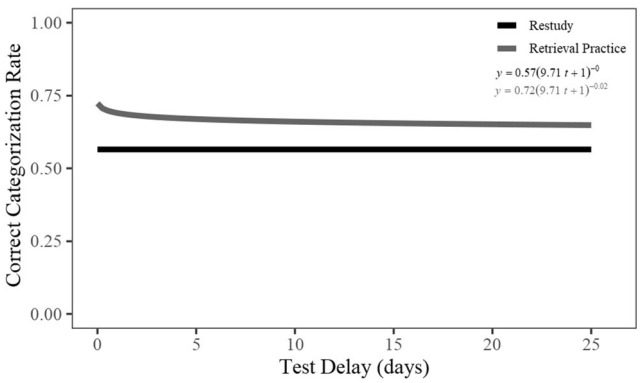

Fig. 6 Forgetting functions fit to categorization data based on median parameter estimates for each study condition (Experiment 2)

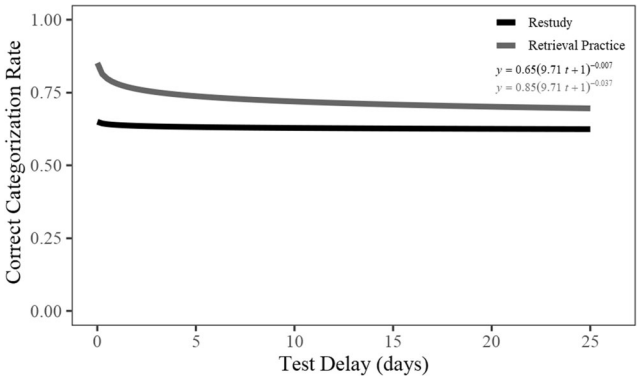

Fig. 7 Forgetting functions fit to the studied set of the categorization data based on median parameter estimates for each study condition (Experiment 2)

For categorization, little forgetting is evident in our data and so may not provide a strong test of whether these conditions lead to differential forgetting (cf. Carpenter et al., 2008; Wheeler et al., 2003). This may be in part because in these set of experiments each final test acts as another learning event for the categories. Even though images were not repeated across the tests, each test assessed knowledge for all the categories. Thus, this design may have traded off natural forgetting of category knowledge with enhanced learning from each of the tests (though those tests did not provide feedback and thus may not have actually served as potent learning events).

General discussion

The goals of these experiments were to (1) replicate previous findings that retrieval practice during study can benefit memory and support transfer of knowledge, and (2) determine whether retrieval practice during study reduces forgetting of rote and categorical knowledge. Overall, Experiments 1 and 2 support the claim that retrieval practice is beneficial to memory and to generalization. Retrieval practice led to better recognition memory and categorization performance than

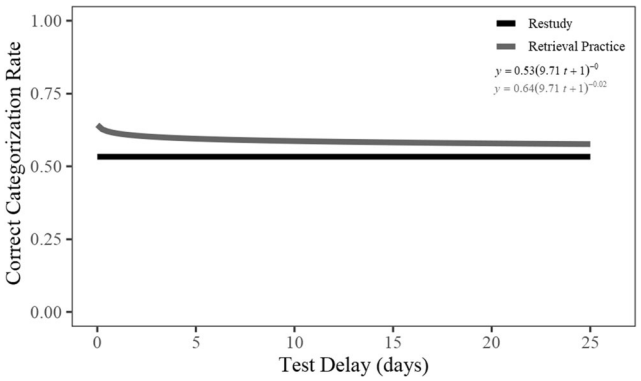

Fig. 8 Forgetting functions fit to the novel set of the categorization data based on median parameter estimates for each study condition (Experiment 2)

restudy in both experiments. Categorization was superior for items from previously retrieved families both for material that was previously studied and for material that was new at time of test. This pattern persisted across delays of up to 25 days.

Differences in forgetting rates were not apparent. The benefits of retrieving on rote knowledge of individual birds and categorical knowledge of bird families were substantial, apparent throughout, but neither decreased nor increased in magnitude when measured by forgetting functions that are known to confirm to the shape of forgetting (Wickelgren, 1974; Wickens, 1998; Wixted, 2004). This result is inconsistent with some of the previous research on testing and forgetting, and this may be due to several reasons.

There are considerable differences between the experiments reported here and previous work that may have affected overall forgetting in different ways. In the current experiments, though no *item* was ever repeated across tests, every *category* was. As a consequence of this design choice, each final test may have acted as a learning event that may have artificially slowed forgetting, which in turn may have clouded the ability to detect differential forgetting across conditions. It may also be that any differences in forgetting rates do not appear in categorization tasks such as the one used here.

Another possible explanation for the disagreement in findings concerns the methods for measuring forgetting. As discussed earlier, previous studies may not have captured forgetting as precisely as we have endeavored to do here. For instance, if a study aims to measure forgetting it must assess memory at multiple time points and then fit that data to a function known to categorize forgetting well.

We maintain that testing as a means of study remains a hugely beneficial strategy, but this strategy may not differentially affect the rate information is forgotten. That is not to say that differential effects on forgetting do not exist, but that they are not clear under these particular circumstances. Of course, further research should be conducted to try to answer these questions more clearly.

The larger aim of these experiments was to evaluate the durability and flexibility of the testing effect. Testing appears to be quite effective on both dimensions: It promotes long-lasting learning that can be generalized beyond the specific materials under study. These results indicate that testing is a powerful study strategy with much potential application in the classroom and beyond.

Author note The authors would like to thank the members of the Human Memory and Cognition Lab for their useful feedback, and Rhea Mundle for her assistance with data collection.

Materials and data from the two experiments can be found on our OSF project page (https://osf.io/suqcx/).

References

Anderson, R. B. (2001). The power law as an emergent property. *Memory & Cognition*, *29*(7), 1061–1068.

Benjamin, A. S., & Pashler, H. (2015). The value of standardized testing a perspective from cognitive psychology. *Policy Insights From the Behavioral and Brain Sciences*, *2*(1), 13–23.

Benjamin, A. S., Bjork, R. A., & Schwartz, B. L. (1998). The mismeasure of memory: When retrieval fluency is misleading as a metamnemonic index. *Journal of Experimental Psychology: General*, *127*(1), 55–68.

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

Butler, A. C. (2010). Repeated testing produces superior transfer of learning relative to repeated studying. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *36*(5), 1118–1133.

Carpenter, S. K., Pashler, H., Wixted, J. T., & Vul, E. (2008). The effects of tests on learning and forgetting. *Memory & Cognition*, *36*(2), 438–448.

Chan, J. C. (2010). Long-term effects of testing on the recall of nontested materials. *Memory*, *18*(1), 49–57.

Estes, W. K. (1956). The problem of inference from curves based on group data. *Psychological Bulletin*, *53*(2), 134–140.

Goldsmith, M., & Koriat, A. (2007). The strategic regulation of memory accuracy and informativeness. *Psychology of Learning and Motivation*, *48*, 1–60.

Jacoby, L. L., & Dallas, M. (1981). On the relationship between autobiographical memory and perceptual learning. *Journal of Experimental Psychology: General*, *110*(3), 306–340.

Jacoby, L. L., Wahlheim, C. N., & Coane, J. H. (2010). Test-enhanced learning of natural concepts: Effects on recognition memory, classification, and metacognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *36*(6), 1441–1451.

Kang, S. H., McDaniel, M. A., & Pashler, H. (2011). Effects of testing on learning of functions. *Psychonomic Bulletin & Review*, *18*(5), 998–1005.

Karpicke, J. D. (2012). Retrieval-based learning active retrieval promotes meaningful learning. *Current Directions in Psychological Science*, *21*(3), 157–163.

Karpicke, J. D., Butler, A. C., & Roediger, H. L., III. (2009). Metacognitive strategies in student learning: Do students practise retrieval when they study on their own?. *Memory*, *17*(4), 471–479.

Kromann, C. B., Jensen, M. L., & Ringsted, C. (2009). The effect of testing on skills learning. *Medical Education*, *43*(1), 21–27.

Kyllonen, P. C., & Tirre, W. C. (1988). Individual differences in associative learning and forgetting. *Intelligence*, *12*(4), 393–421.

Loftus, E. F. (2003). Make-believe memories. *American Psychologist*, *58*(11), 867–873.

Lyle, K. B., & Crawford, N. A. (2011). Retrieving essential material at the end of lectures improves performance on statistics exams. *Teaching of Psychology*, *38*(2), 94–97.

MacLeod, C. M., Gopie, N., Hourihan, K. L., Neary, K. R., & Ozubko, J. D. (2010). The production effect: Delineation of a phenomenon. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *36*(3), 671–685.

Morey, R. D., & Rouder, J. N. (2015). BayesFactor: Computation of Bayes factors for common designs (R Package Version 0.9.12-2) [Computer software]. Retrieved from https://CRAN.R-project.org/package=BayesFactor

Morris, C. D., Bransford, J. D., & Franks, J. J. (1977). Levels of processing versus transfer appropriate processing. *Journal of Memory and Language*, *16*(5), 519–533.

Murre, J. M., & Chessa, A. G. (2011). Power laws from individual differences in learning and forgetting: Mathematical analyses. *Psychonomic Bulletin & Review*, *18*(3), 592–597.

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

Roediger, H. L., & Karpicke, J. D. (2006a). Test-enhanced learning taking memory tests improves long-term retention. *Psychological Science*, *17*(3), 249–255.

Roediger, H. L., & Karpicke, J. D. (2006b). The power of testing memory: Basic research and implications for educational practice. *Perspectives on Psychological Science*, *1*(3), 181–210.

Roediger, H. L., Agarwal, P. K., McDaniel, M. A., & McDermott, K. B. (2011a). Test-enhanced learning in the classroom: Long-term improvements from quizzing. *Journal of Experimental Psychology: Applied*, *17*(4), 382–395.

Roediger, H. L., Putnam, A. L., & Smith, M. A. (2011b). Ten benefits of testing and their applications to educational practice. *Psychology of Learning and Motivation, 55*, 1–36.

Rohrer, D., Taylor, K., & Sholar, B. (2010). Tests enhance the transfer of learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *36*(1), 233–239.

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

Rubin, D. C., & Wenzel, A. E. (1996). One hundred years of forgetting: A quantitative description of retention. *Psychological Review*, *103*(4), 734–760.

Runquist, W. N. (1983). Some effects of remembering on forgetting. *Memory & Cognition*, *11*(6), 641–650.

Schmidt, R. A., & Bjork, R. A. (1992). New conceptualizations of practice: Common principles in three paradigms suggest new concepts for training. *Psychological Science*, *3*(4), 207–217.

Steyvers, M., & Benjamin, A. S. (2018). The joint contribution of participation and performance to learning functions: Exploring the effects of age in large-scale data sets. *Behavior Research Methods*, 1–13. Advance online publication. doi:https://doi.org/10.3758/s13428-018-1128-2

The 47th PDK/Gallup Poll of the public's attitudes toward the public schools: Testing doesn't measure up for Americans. (2015). *Phi Delta Kappan*, *97*(1). doi:https://doi.org/10.1177/0031721715602231

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

Toppino, T. C., & Cohen, M. S. (2009). The testing effect and the retention interval: Questions and answers. *Experimental Psychology*, *56*(4), 252–257.

Tullis, J. G., Finley, J. R., & Benjamin, A. S. (2013). Metacognition of the testing effect: Guiding learners to predict the benefits of retrieval. *Memory & Cognition*, *41*(3), 429–442.

Tullis, J. G., Fiechter, J. L., & Benjamin, A. S. (2018). The efficacy of learners' testing choices. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 44*(4), 540–552.

Wheeler, M. A., & Roediger, H. L. (1992). Disparate effects of repeated testing: Reconciling Ballard's (1913) and Bartlett's (1932) results. *Psychological Science*, *3*(4), 240–245.

Wheeler, M., Ewers, M., & Buonanno, J. (2003). Different rates of forgetting following study versus test trials. *Memory*, *11*(6), 571–580.

Wickelgren, W. A. (1974). Single-trace fragility theory of memory dynamics. *Memory & Cognition*, *2*(4), 775–780.

Wickens, T. D. (1998). On the form of the retention function: Comment on Rubin and Wenzel (1996): A quantitative description of retention. *Psychological Review, 105*(2), 379–386.

Wiklund-Hörnqvist, C., Jonsson, B., & Nyberg, L. (2014). Strengthening concept learning by repeated testing. *Scandinavian Journal of Psychology*, *55*(1), 10–16.

Wixted, J. T. (2004). On common ground: Jost's (1897) law of forgetting and Ribot's (1881) law of retrograde amnesia. *Psychological Review*, *111*(4), 864–879.