Mem Cogn (2017) 45:81–92
DOI 10.3758/s13421-016-0641-8

Comparing the testing effect under blocked and mixed practice: The mnemonic benefits of retrieval practice are not affected by practice format

Magdalena Abel[1] · **Henry L. Roediger III**[1]

Published online: 27 July 2016
© Psychonomic Society, Inc. 2016

Abstract The act of retrieving information modifies memory in critical ways. In particular, testing-effect studies have demonstrated that retrieval practice (compared to restudy or to no testing) benefits long-term retention and protects from retroactive interference. Although such testing effects have previously been demonstrated in both between- and within-subjects manipulations of retrieval practice, it is less clear whether one or the other testing format is most beneficial on a final test. In two paired-associate learning experiments conducted under typical testing-effect conditions, we manipulated restudy and test trials using either blocked or mixed practice conditions while equating other factors. Retrieval-practice and restudy trials were presented either separately in different blocks (blocked practice) or randomly intermixed (mixed practice). In Experiment 1, recall was assessed after short and long delay intervals; in Experiment 2, the final memory test occurred after a short delay, but with or without an interfering activity before the final test. In both experiments, typical testing effects emerged, and critically, they were found to be unaffected by practice format. These results support the conclusion that testing effects are robust and emerge to equal extents in both blocked and mixed designs. The generality of testing effects further encourages the application of retrieval practice as a memory enhancer in a variety of contexts, including education.

Keywords Retrieval practice · Testing effect · Delay · Interference · Practice format

✉ Magdalena Abel
magdalena.abel@ur.de

[1] Department of Psychology, Washington University in St. Louis, One Brookings Drive, Box 1125, St. Louis, MO 63130-4899, USA

The general assumption in education, and often in cognitive psychology, is that the act of testing memory is a neutral affair; tests are given to assess one's knowledge, not to change it. However, 40 years ago R.A. Bjork (1975) argued that recall not only measures memory but modifies it in several ways, often positively. In recent years, the testing effect—the benefit of the act of recall on later retention—has been frequently studied, and much has been learned. The *testing effect* refers to the finding that practicing the retrieval of previously studied material can boost memory and enhance long-term retention for the tested materials in comparison to a no-exposure control condition (e.g., Wheeler & Roediger, 1992), or even in comparison to control conditions in which the materials are reread for the same amount of time (e.g., Carrier & Pashler, 1992; Roediger & Karpicke, 2006; for a review, see McDermott, Arnold, & Nelson, 2014). Moreover, retrieval has been shown to protect the practiced material from the detrimental influence of subsequent learning (i.e., from *retroactive interference*; Halamish & Bjork, 2011; Potts & Shanks, 2012). On the other hand, research on retrieval-induced forgetting indicates that retrieval can also entail negative effects for memory of items related to the retrieved items when retrieval is carried out only selectively (Anderson, Bjork, & Bjork, 1994; see also Anderson, 2003). When retrieval practice occurs for some but not all information, such selective retrieval can cause forgetting of related, but unpracticed contents (relative to a control condition without any practice; for a recent meta-analysis and review on retrieval-induced forgetting, see Murayama, Miyatsu, Buchli, & Storm, 2014). Although the present experiments are concerned with the testing effect, their main motivation was derived from prior work on retrieval-induced forgetting. In addition, the present work also relates to the generation effect, so we briefly review relevant studies from these other domains before providing the rationale for our experiments.

The special importance of retrieval as a memory modifier is underscored by studies investigating whether the observed effects of retrieval practice are really specific to retrieval, or might alternatively also arise when memories are strengthened in a different way instead (e.g., by means of restudying). In particular, several studies have documented the point that retrieval-induced forgetting emerges only after retrieval practice, but not after restudy, and of course (by definition) the retrieval-practice effect occurs after testing (and the comparison condition is often a restudy control). Importantly, for both effects, such differences between retrieval practice and restudy have been successfully demonstrated with between-subjects *and* within-subjects manipulations of the two practice types. For instance, the positive influence of retrieval practice on long-term retention in the testing-effect literature has been demonstrated when different subjects engaged in retrieval practice or restudy (e.g., Karpicke & Roediger, 2008; Pyc & Rawson, 2010), but also when the same subjects practiced some materials by means of retrieval practice and others by means of restudy (e.g., Butler, 2010; Zaromb & Roediger, 2010). Similarly, the negative influence of selective retrieval practice for related but unpracticed material has also been shown to occur when different subjects (e.g., Anderson, Bjork, & Bjork, 2000; Bäuml & Aslan, 2004; Ciranni & Shimamura, 1999) or the same subjects (e.g., Hanslmayr, Staudigl, Aslan, & Bäuml, 2010; Wimber, Rutschmann, Greenlee, & Bäuml, 2009) engaged in selective retrieval practice and/or restudy.

Yet another important issue has been less frequently examined: Few studies have investigated whether the format of the practice type matters: Are the effects of retrieval practice affected by whether practice occurs in blocks of trials (e.g., blocks of items are restudied or are tested) or, rather, the two types of trials are mixed together? A recent study by Dobler and Bäuml (2013) was the first to address whether practice format was of relevance for retrieval-induced forgetting. Their results showed that when practice was blocked and retrieval-practice and restudy trials were clearly separated, only selective retrieval practice (but not restudy) led to retrieval-induced forgetting (thus replicating prior work; e.g., Hanslmayr et al., 2010). However, when retrieval-practice and restudy trials were randomly intermixed, retrieval-induced forgetting emerged after both types of practice. Dobler and Bäuml interpreted their results as a reflection of dynamic effects between retrieval practice and restudying when the two types of trials are intermixed. That is, in relating their findings to previous work from the task-switching literature (e.g., Allport, Styles, & Hsieh, 1994; Campbell, 2005; Meuter & Allport, 1999), Dobler and Bäuml argued that, when switching from more effortful retrieval-practice trials to the easier restudy trials, subjects might still engage in retrieval practice to some degree even on restudy trials, thereby causing retrieval-induced forgetting after both retrieval-practice and restudy trials. Thus, it is not restudying that causes forgetting, but the fact that, under certain conditions, subjects can be led to retrieve during restudy (for related work, see also Jacoby & Wahlheim, 2013; Tullis, Benjamin, & Ross, 2014).

On the basis of this reasoning, the question arises whether such dynamic effects between retrieval-practice and restudy trials occurring with mixed practice in retrieval-induced forgetting studies may also be of relevance for the testing effect. The case seems to be slightly different, though, and the consequences of mixed practice might not directly translate from one paradigm to the other. Indeed, several previous studies on the testing effect have applied randomly intermixed restudy and retrieval-practice trials, and these studies nevertheless revealed intact testing effects (e.g., Carpenter & DeLosh, 2006; Carpenter, Pashler, Wixted, & Vul, 2008; Karpicke & Zaromb, 2010; see also Rowland, 2014, for a recent meta-analysis). Thus, mixed practice is unlikely to affect the testing effect in the same "all-or-none" way as it appears to do in retrieval-induced forgetting research; it remains unclear, though, whether practice format is relevant for the size of the testing effect. Dobler and Bäuml (2013) suggested that, during mixed practice, subjects might keep engaging in retrieval practice when switching from the harder (retrieval-practice) trials to the easier (restudy) trials. However, this particular type of switching might only occur for a part of the to-be-restudied materials (perhaps later in practice), so that differences between blocked and mixed practice might be more subtle with regard to the benefits of retrieval practice.

Interestingly, a series of recent studies addressed a related topic. In particular, these experiments were specifically designed to examine a potential parallel between testing effects (i.e., retrieval from episodic memory) and generation effects (i.e., retrieval from semantic memory). The *generation effect* refers to the finding that items generated from semantic memory are remembered better on a later test than items that were simply read, and prior work has shown that this effect can vary with list composition and is, at least under certain conditions, larger with mixed lists (containing randomly mixed read *and* generate trials) than with pure lists (containing read *or* generate trials; e.g., Nairne, Riegler, & Serra, 1991; Slamecka & Katsaiti, 1987; see McDaniel & Bugg, 2008, for a review). Three recent studies investigated whether the testing effect, in parallel to the generation effect, was also sensitive to list composition. In one of these studies, Rowland, Littrell-Baez, Sensenig, and DeLosh (2014) reported no influence of list composition on the testing effect, whereas, in another study, Mulligan and Peterson (2015) found that the testing effect behaved similarly to the generation effect and was larger with mixed than with pure lists of retrieval-practice and restudy trials. In a follow-up study, Mulligan, Susser, and Smith (2016) replicated the original finding reported by Mulligan and Peterson under varying

conditions and confirmed across four experiments that the testing effect can be modulated by list composition in a way similar as the generation effect—as long as specific experimental procedures are applied, ones under which the generation effect is also sensitive to list composition. For example, in accordance with theoretical accounts of list-composition effects (e.g., the item-order account; see Nairne et al., 1991), such effects may hinge on the exact nature of the final test. For example, they may only emerge when final free-recall tests are conducted soon after study of each single (pure or mixed) list, thus creating unique retrieval sets for each pure list that aid recall, relative to mixed lists (for details, see Mulligan et al., 2016; see also our General Discussion).

The results just reviewed have important implications for theoretical accounts of the testing effect, because they suggest that testing and generation effects might be caused by similar mechanisms, at least in some situations (but see Karpicke & Zaromb, 2010). Yet, because the three previous studies on the role of list composition (Mulligan & Peterson, 2015; Mulligan et al., 2016; Rowland et al., 2014) had purposefully applied conditions that are conducive for list-composition effects for the generation effect, their results may not apply to the issue of whether the testing effect can be modulated by practice format when experimental conditions typical for testing-effect studies are used (e.g., when the final test comprises both restudied and retrieval practiced items and is conducted not only after short but also after prolonged delays, or more generally under conditions that make recall more difficult).

On the basis of the results reported by Dobler and Bäuml (2013), a contrasting prediction on the role of practice format (blocked vs. randomly intermixed) for the testing effect arises under more standard conditions used in testing-effect studies. In particular, we predicted from their work that typical testing effects would be larger with blocked than with mixed practice lists, because in the latter case subjects would covertly retrieve on the restudy trials. This prediction assumes that the dynamic effects between retrieval-practice and restudy trials with mixed practice would generalize from studies of retrieval-induced forgetting to the testing effect with a more standard design. As noted, each of the recent prior studies on the role of list composition was modeled after research on the generation effect and based on procedures that differ from those typically used for testing-effect studies. One important difference is that none of the prior studies used retention intervals of more than a few minutes or, more generally, created conditions with increased difficulty at final test, even though the testing effect is often larger on delayed or difficult tests (e.g., Halamish & Bjork, 2011; Roediger & Karpicke, 2006; see also Rowland, 2014).

The goal of the present study was to use typical testing-effect conditions to explore whether testing effects may be modulated by practice format in a different way than suggested by the previous studies on the role of list composition. In Experiment 1, subjects studied Swahili–English vocabulary pairs, engaged in repeated retrieval-practice cycles (plus feedback) and restudy cycles, and did so in either a blocked or a mixed fashion. A final memory test was given after 5 min or 1 week, thus increasing the difficulty at test. In Experiment 2, subjects studied unrelated word pairs, engaged repeatedly in either blocked or mixed practice, and completed a final test in the presence or absence of retroactive interference, which constitutes another way of manipulating difficulty at test. Our experiments will show whether Dobler and Bäuml's (2013) findings on retrieval-induced forgetting generalize to the testing effect by examining whether study trials have greater impact (thus reducing the testing effect) during random than during blocked practice. That is, mixed practice should reduce the testing effect, because random switches between practice types might encourage subjects to keep engaging in retrieval practice during restudy trials. Moreover, the present experiments will also show whether such a pattern depends on the difficulty of the final test, or whether the results are comparable across short and long delays (in Exp. 1) and conditions with and without retroactive interference (in Exp. 2).

Experiment 1

Method

Subjects

One hundred four undergraduates at Washington University in St. Louis were recruited for the study. Six of the subjects did not return for the second session of the experiment, and two further subjects were excluded because they notified the experimenter about having prior knowledge of Swahili. The final sample thus included 96 subjects that were evenly distributed across the four conditions (i.e., $n = 24$ in each condition). The mean age was 20.2 years ($SD = 2.4$ years); there were 14 female and ten male subjects in each of the two short-delay conditions, and 15 female and nine male subjects in each of the two long-delay conditions. Subjects were tested either singly or in small groups, and they received course credit or $10 for completing the study.

Material

Twenty four Swahili–English word pairs were selected from the norms provided by Nelson and Dunlosky (1994), and were divided into two sets of 12 pairs. The two sets occurred equally often in the restudy and retrieval-practice conditions across subjects.

Design

The experiment employed a 2 × 2 × 2 mixed-factorial design. All subjects initially studied the 24 pairs under the same presentation conditions. The first factor, Practice (retrieval practice or restudy), was manipulated within subjects; after the initial study phase, all subjects were asked to practice retrieval for one half of the pairs and to restudy the other half. The second factor, Practice Format (blocked practice, mixed practice), was manipulated between subjects. Forty-eight subjects engaged in blocked practice: In this condition, restudy and retrieval practice were carried out in clearly separable blocks, with the sequence of restudy and retrieval-practice blocks being counterbalanced across subjects. For instance, if practice started with restudy, then all 12 word pairs from one of the two sets of material were repeatedly restudied without any retrieval-practice trials intermixed; subsequently, subjects engaged in a block of retrieval-practice cycles for the remaining 12 word pairs. The other 48 subjects engaged in mixed practice: Here, restudy and retrieval-practice trials were randomly intermixed, and the sequence of restudy and retrieval-practice trials was not predictable. To avoid differences in spacing across the blocked- and mixed-practice conditions (see Cepeda, Pashler, Vul, Wixted, & Rohrer, 2006), the subjects in the mixed-practice condition also completed two phases of practice, with half of the word pairs being practiced during the first phase and the other half being practiced during the second phase. In contrast to the blocked-practice condition, however, these two practice phases in the mixed-practice condition always entailed both retrieval-practice and restudy trials, intermixed randomly. The last factor was Delay (5 min, 7 days), manipulated between subjects. A final recall test on all 24 word pairs was conducted after either 5 min or 7 days, as in previous studies on the testing effect (e.g., Roediger & Karpicke, 2006).

Procedure

Study phase In an initial study phase, subjects were presented with 24 Swahili–English vocabulary pairs under intentional learning conditions. Word pairs were presented in a random sequence, for 4 s each, centrally on a computer screen.

Practice phase After initial study, the subjects were informed that all word pairs would be practiced again in two separate phases (with half of the pairs being practiced during the first phase and the other half during the second phase). Note that the two separate phases were also introduced in the mixed-practice condition, to avoid differences in spacing/practice lag relative to the blocked-practice condition, as we described in the Design section. In the blocked-practice condition, one of the practice phases contained only restudy trials, whereas the other phase contained only retrieval-practice trials. In contrast, in the mixed-practice condition, both practice phases contained randomly intermixed retrieval-practice and restudy trials (with half of the word pairs during each phase being assigned to retrieval practice, and half to restudy). There were no further differences between the two practice format conditions. On restudy trials, word pairs were presented in intact form for 7 s each, and subjects were asked to type in the English translation of the Swahili words. On retrieval-practice trials, the Swahili words were presented for 5 s, and subjects were asked to type in the English translation during this 5-s interval if they could remember it. After 5 s, the correct answer was presented for an additional 2 s. During each practice phase, all 12 word pairs were practiced three times in the same manner, with a random sequence of pairs on each of the three practice cycles. When practice was completed, all subjects were asked to solve simple arithmetic equations for 5 min.

Final-test phase The subjects in the short-delay condition completed the final test after the 5-min distractor task; the subjects in the long-delay condition left the lab and returned to take the same test after 7 days. On the final test, all 24 word pairs were tested: The Swahili words were presented in random order and for 10 s each, with subjects being instructed to type in the English translations of the words. After completing the test, subjects were debriefed and thanked for their participation.

Results

Performance on retrieval-practice cycles

Figure 1a shows mean recall on the three repeated practice cycles that contained retrieval-practice trials. A 3 × 2 × 2 analysis of variance (ANOVA) with the within-subjects factor Retrieval-Practice Cycle (first, second, third) and the between-subjects factors Practice Format (blocked, mixed) and Delay (5 min, 7 days) revealed a significant main effect of retrieval-practice cycle, $F(2, 184) = 449.98$, $MSE = 160.73$, $p < .001$, $\eta^2 = .83$. Retrieval practice plus feedback enhanced recall from the first retrieval-practice cycle to the second (15.8 % vs. 50.2 %), $t(95) = 18.58$, $p < .001$, $d = 1.88$, and from the second retrieval-practice cycle to the third (50.2 % vs. 70.1 %), $t(95) = 13.80$, $p < .001$, $d = 1.40$. No other main effects or interactions reached significance, all Fs < 1.0, indicating the general equivalence of the various between-subjects conditions.

Retention on the final test

Figure 1b displays mean recall on the final test. A 2 × 2 × 2 ANOVA showed a significant main effect of type of practice, $F(1, 92) = 53.72$, $MSE = 178.91$, $p < .001$, $\eta^2 = .37$. Recall was superior after retrieval practice as compared to restudy, both

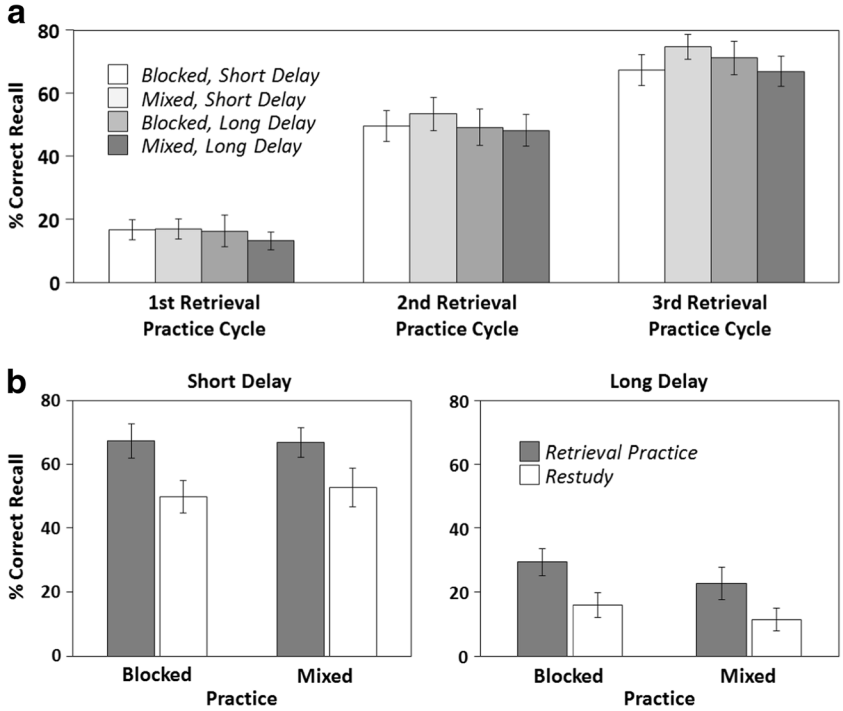

Fig. 1 (a) Mean recall performance on the first, second, and third retrieval-practice cycles, shown as a function of practice-format conditions (blocked, mixed) and delay conditions (short, long). (b) Mean recall performance on the final test, shown separately for the short- and long-delay conditions and plotted as a function of both practice (retrieval practice, restudy) and practice format (blocked, mixed). Error bars represent ±1 standard error

after the 5-min delay (67.2 % vs. 51.4 %), $t(47) = 5.54$, $p < .001$, $d = 0.80$, and after the 7-day delay (26.2 % vs. 13.7 %), $t(47) = 4.89$, $p < .001$, $d = 0.71$. There was also a significant main effect of delay, $F(1, 92) = 53.72$, $MSE = 78.98$, $p < .001$, $\eta^2 = .46$, because recall declined across the 7-day delay, irrespective of whether word pairs had been subject to retrieval practice (67.2 % vs. 26.2 %), $t(94) = 8.48$, $p < .001$, $d = 1.73$, or restudy (51.4 % vs. 13.7 %), $t(94) = 7.92$, $p < .001$, $d = 1.65$. Most importantly, neither the main effect of practice format nor any interactions reached significance, all Fs < 1.0, which indicates that varying retrieval practice and restudy in blocks relative to randomly intermixing the two types of trials did not significantly alter the pattern of results at either retention interval.

A power analysis using G*Power 3 (Faul, Erdfelder, Lang, & Buchner, 2007) showed that the effective power to detect even a small-sized interaction effect ($f = 0.10$; Cohen, 1988) with the present sample size and at an alpha-level of .05 was .75. Moreover, because null hypothesis significance testing cannot provide support for null hypotheses (see Gallistel, 2009; Wagenmakers, 2007), we followed Masson's (2011) guidelines and used the Bayesian information criterion (BIC) to compute posterior probabilities. In general, this approach assumes that two competing hypotheses are equally likely a priori, before data collection, and generates posterior probabilities for the hypotheses being correct given a set of observed data. Here, we used the approach to generate such posterior probabilities for H_0 and H_1, given the data (D) collected in Experiment 1. First, we calculated difference scores to capture the magnitude of the testing effect (i.e., we subtracted the final recall performance after restudy from the final recall performance after retrieval practice). Then, two separate analyses were run for the short- and long-delay conditions on the basis of these difference scores. In the short-delay conditions, the resulting posterior probabilities were $P_{BIC}(H_0 \mid D) = .856$ and $P_{BIC}(H_1 \mid D) = .144$. Similar probabilities were obtained for the long-delay conditions, with $P_{BIC}(H_0 \mid D) = .864$ and $P_{BIC}(H_1 \mid D) = .136$. Following Raftery (1995; see also Masson, 2011), this outcome can be interpreted as positive evidence in favor of the null hypothesis (i.e., practice format does not affect the size of the testing effect with the present experimental variables).

Discussion

The results of Experiment 1 are generally consistent with the literature on the testing effect: Retrieval practice led to enhanced retention on the final test in comparison to a restudy control condition. This retrieval-specific boost in final-test performance was similar in size after 5 min and 7 days. While some previous studies found the testing effect to increase or only to occur with long retention intervals when tests did not provide feedback (e.g., Bäuml, Holterman, & Abel, 2014; Congleton & Rajaram, 2012; Roediger & Karpicke,

2006), the present experiment is consistent with other reports of similar effects after a short and a long delay when feedback is provided on the initial tests (e.g., Carpenter et al., 2008). The data reported by Kornell, Bjork and Garcia (2011, Exp. 2), moreover, support the idea that provision of feedback after retrieval attempts may be a crucial difference between studies that show immediate testing effects and those that do not (like Roediger & Karpicke, 2006, and others). Corrective feedback was also provided in the present study, and the results are consistent with those reported by Kornell et al. (2011).

Most importantly, however, the data of Experiment 1 show that the size of the testing effect was not affected by whether subjects were asked to engage in retrieval-practice and restudy trials in a blocked or in a mixed fashion at either retention interval. This result indicates that the prior findings and conclusions by Dobler and Bäuml (2013) may not generalize to the testing effect; the testing effect, in contrast to retrieval-induced forgetting, does not seem to be affected by practice format. In particular, the fact that the testing effect was not sensitive to practice format under standard experimental procedures used to study the effect and that it occurred after both short (5-min) and long (7-day) retention intervals indicates that retrieval-practice effects are robust across these manipulations; in particular, the findings show that the testing effect is unaffected by practice format, irrespective of whether the final test is relatively easy (after short delay) or relatively difficult (after prolonged delay). Experiment 2 was designed to conceptually replicate these effects.

Experiment 2

Experiment 2 was conducted to replicate the null effects of the blocked-versus-mixed manipulation of restudy and retrieval practice conditions and to determine whether the results of Experiment 1 would generalize to conditions in which recall was assessed with and without retroactive interference, rather than after short and long retention intervals (i.e., as another way of establishing relatively easy or difficult testing conditions). Halamish and Bjork (2011) showed that retrieval practice in comparison to restudy can reduce the susceptibility to retroactive interference (for similar results, see Potts & Shanks, 2012). In Experiment 2, we examined the role of practice format (blocked or mixed) in the testing effect across this variable.

Method

Subjects

Ninety four undergraduates at Washington University in St. Louis participated in the study and were compensated with course credit or $10. Eight subjects did not engage in practice (i.e., they did not type in their answers or at least the answers were not recorded by the computer), and they were therefore excluded from the data set. Thus, 86 subjects remained, with 44 randomly assigned to the blocked-practice condition (34 female, 10 male subjects) and 42 to the mixed-practice condition (28 female, 14 male subjects). Their mean age was 20.4 years (*SD* = 4.6 years). Subjects were tested individually or in small groups.

Material

Forty eight unrelated word pairs were created for paired-associate learning from 96 single words selected from norms provided by Van Overschelde, Rawson, and Dunlosky (2004). The words were taken from different semantic categories (e.g., the words "lettuce" and "sandal") and were randomly turned into paired associates (e.g., "lettuce–sandal"). Subjects were asked to complete two sessions of the same task, with the two differing only in whether or not retroactive interference was introduced before the final test. We used 24 paired associates for the first session, and the remaining 24 for the second (with the sets of material being counterbalanced across the two sessions). As in Experiment 1, the two sets of materials were randomly split into two further subsets of 12 paired associates, and assignment of subsets to retrieval-practice and restudy conditions was counterbalanced across subjects. The use of a new set of materials was intended to generalize our findings across this dimension, as well as to more easily instantiate the variable of interference.

To induce retroactive interference, 24 additional single items were chosen from the Van Overschelde et al. (2004) norms (e.g., the word "ring" to be paired with "lettuce"). These 24 items were presented as new response terms to the previously studied stimulus terms to induce A–B, A–D retroactive interference; during selection, care was taken to insure that old and new response terms never began with the same initial letters. For instance, if the word pair "lettuce–sandal" had been studied initially, the word pair "lettuce–ring" might be presented for additional study.

Design

The experiment employed a 2 × 2 × 2 mixed-factorial design. The first factor, Practice (retrieval practice, restudy), was manipulated within subjects. After initial study of all items, subjects were asked to restudy one half of the material and to practice retrieval of the other half. The second factor, Practice Format (blocked, mixed), was again manipulated between subjects. As in Experiment 1, half of the subjects engaged in blocked practice, whereas the other half engaged in mixed practice (using the same strictures as in Exp. 1). The last factor, Retroactive Interference (no interference, interference), was manipulated within subjects. All subjects were asked to engage in two sessions of the same task (including

an initial study phase, a practice phase with either blocked or mixed practice, and a final test). The two sessions differed only in whether or not retroactive interference was induced after study and before the final test. Therefore, in one of the two sessions, subjects were asked to complete an unrelated distractor task between practice and the final test (nonspecific interference, the control). In the other session, after restudy and retrieval practice, subjects studied additional paired associates with old stimulus terms but new response terms (inducing A–B, A–D retroactive interference). To reiterate, this was done to investigate whether retrieval practice (relative to restudy) would reduce susceptibility to retroactive interference (Halamish & Bjork, 2011). Critically, by manipulating practice format we can determine whether the pattern of results would be differently affected by blocked and mixed practice.

Procedure

Study phase Each of the two sessions began with an initial study phase, and 24 paired associates (e.g., "lettuce–sandal") were presented in random order, for 4 s each, in the center of a computer screen. Subjects were asked to try to memorize the word pairs for a later test.

Practice phase After initial study, subjects were informed that all word pairs would be practiced again in two separate practice phases (as in Exp. 1, half of all word pairs were practiced during the first practice phase, and the other half during the second practice phase). In the blocked-practice condition, one phase contained only restudy trials, whereas the other phase contained only retrieval-practice trials. In contrast, in the mixed-practice condition, both practice phases contained randomly intermixed retrieval-practice and restudy trials (with half of the word pairs during each phase being assigned to retrieval practice, the other half being assigned to restudy). As in Experiment 1, on restudy trials, the paired associates were presented intact for 7 s each (e.g., "lettuce–sandal"), and subjects were asked to type in each response term (e.g., "sandal"). On retrieval-practice trials, the stimulus terms and the matching initial letters of the response terms were presented for 5 s each (i.e., "lettuce–s__?"), and subjects were asked to recall and type in the full response term during this interval. After 5 s, the correct answer was presented for an additional 2 s (i.e., "lettuce–sandal"). As in Experiment 1, all paired associates were practiced three times in the same manner, with word pairs being presented in random sequence on each of the three practice cycles.

Induction of retroactive interference As we described above, there were two study lists and tests in the experiment, and the corresponding sessions differed only in whether or not A–B, A–D retroactive interference was induced before the final test. Whether retroactive interference occurred in the first or the second session was counterbalanced across subjects. In the baseline condition, no specific retroactive interference was induced, and after practice subjects were asked to work on an unrelated distractor task for 5 min (i.e., to write down as many American presidents as possible). However, in the experimental condition with retroactive interference, subjects were instead asked to study 24 new paired associates, with new response terms to the previously studied stimulus terms (e.g., "lettuce–ring"). The new paired associates were again presented for 4 s each and in random order. Three consecutive study cycles were given in this manner, to ensure that subjects would try to memorize the new response terms.

Final-test phase Each of the two sessions of the experiment ended with a final memory test. The test was given right after the distractor task (for the session without retroactive interference) or after additional study of the new response terms (in the session with retroactive interference). The stimulus terms plus the initial letters of the response terms were provided as retrieval cues (e.g., "lettuce–s__?"), in random order and at a rate of 10 s each. Subjects were asked to type in the corresponding response terms. The additionally studied word pairs for the set of materials in which retroactive interference was manipulated were tested in the same manner (e.g., "lettuce–r__?"), but only after all initially studied word pairs had already been tested. When the first session of the experiment was completed, subjects were offered a short break. Afterward, they were asked to complete the second session with a new list. At the end, subjects were debriefed and thanked for their participation.

Results

Performance on retrieval-practice cycles

Figure 2a shows mean recall on the three consecutive retrieval-practice cycles. A 3 × 2 × 2 ANOVA with the within-subjects factors Retrieval-Practice Cycle (first, second, third) and Retroactive Interference (no interference, interference), as well as with the between-subjects factor Practice Format (blocked, mixed) revealed a significant main effect of the factor Retrieval-Practice Cycle, $F(2, 168) = 415.39$, $MSE = 269.72$, $p < .001$, $\eta^2 = .83$. As in Experiment 1, recall increased from the first retrieval-practice cycle to the second (41.1 % vs. 78.1 %), $F(1, 84) = 447.54$, $MSE = 262.82$, $p < .001$, $\eta^2 = .84$, and from the second to the third (78.1 % vs. 90.0 %), $F(1, 84) = 83.65$, $MSE = 147.50$, $p < .001$, $\eta^2 = .50$. There was neither a significant main effect of Interference, $F(1, 84) = 1.96$, $MSE = 399.22$, $p = .165$, $\eta^2 = .02$, nor of Practice Format, $F(1, 84) < 1.0$, which indicates that performance on retrieval-practice cycles did not generally differ between the conditions. However, although no further interactions reached significance (all Fs < 1.0), we did observe a significant interaction between the factors Retrieval-Practice

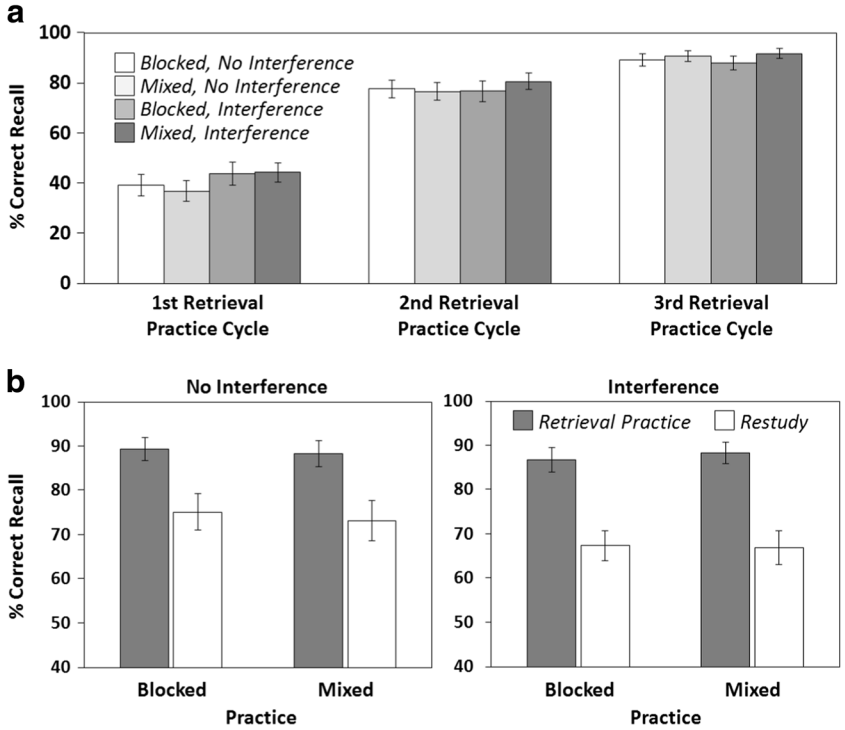

Fig. 2 (a) Mean recall performance on the first, second, and third retrieval-practice cycles, shown as a function of practice-format ?twb=.27w?>conditions (blocked, mixed) and interference conditions (no interference, interference). (b) Mean recall performance on the final test, shown separately for the conditions with and without retroactive interference and plotted as a function of both practice (retrieval practice, restudy) and practice format (blocked, mixed). Error bars represent ±1 standard error

Cycle and Retroactive Interference, $F(2, 168) = 4.41$, $MSE = 100.45$, $p = .020$, $\eta^2 = .05$. Follow-up t-tests demonstrated that recall differed between conditions with and without retroactive interference, but only on the first retrieval-practice cycle (without interference, 38.1 % correct; with interference, 44.1 % correct), $t(85) = 2.36$, $p = .021$, $d = 0.25$. On the second and third retrieval-practice cycles, recall was comparable with and without retroactive interference, all ts < 1.0. Thus, although subjects started the first retrieval-practice cycle with higher recall in the experimental condition in which interference would later occur, because the following retrieval-practice cycles led to similar levels of recall, we deem the difference observed on the first cycle to be spurious. Because results across the four conditions were quite similar on the second and the third sessions of retrieval practice, the spurious result on the first trial can safely be assumed to have no bearing on the results from manipulating retroactive interference, which of course occurred later in the procedure.

Retention on the final test

Figure 2b shows mean recall on the final test. A 2 × 2 × 2 ANOVA revealed a significant main effect of Practice, $F(1, 84) = 89.84$, $MSE = 293.29$, $p < .001$, $\eta^2 = .52$, reflecting the fact that recall was better after retrieval practice than after restudy (88.2 % vs. 70.7 %). In addition, the ANOVA showed a significant main effect of Retroactive Interference, $F(1, 84) = 5.29$, $MSE = 285.22$, $p = .024$, $\eta^2 = .06$, which was accompanied by a significant interaction of the two factors, $F(1, 84) = 6.18$, $MSE = 114.11$, $p = .015$, $\eta^2 = .07$. Although recall was not affected by retroactive interference after retrieval practice (88.9 % vs. 87.5 %), $t(85) < 1.0$, $p = .486$, $d = 0.08$, recall of restudied word pairs was impaired by studying the additional list (74.2 % vs. 67.2 %), $t(85) = 3.03$, $p = .003$, $d = 0.33$. Critically, neither the main effect of practice format nor any further interactions reached significance, all Fs < 1.0, which suggests that, as in Experiment 1, the manipulation of restudy and retrieval practice was the same whether these were manipulated randomly or in blocks.

A power analysis using G*Power 3 (Faul et al., 2007) showed that the effective power to detect a small-sized interaction effect ($f = .10$; Cohen, 1988) with the present sample size and an alpha level of .05 was .64. In parallel to Experiment 1, we used the BIC to compute posterior probabilities for H_0 and H_1, given the collected data (for details, see Masson, 2011). Again, two separate analyses on the difference scores (i.e., final recall performance after retrieval practice minus final recall performance after restudy) were conducted for the conditions with and without interference. In the condition without interference, posterior probabilities were $P_{BIC}(H_0 \mid D) = .901$ and $P_{BIC}(H_1 \mid D) = .099$; in the condition with interference, $P_{BIC}(H_0 \mid D)$ was .892 and $P_{BIC}(H_1 \mid D)$ was

.108. These analyses can again be interpreted as providing positive evidence in favor of the null hypothesis (see Raftery, 1995); under the present experimental conditions, practice format does not affect the magnitude of the testing effect, either with or without retroactive interference.

Discussion

The results of Experiment 2 are again consistent with previous work on the mnemonic benefits of retrieval practice. Although retroactive interference only had a small effect in Experiment 2, retrieval practice was found to shield memories from its detrimental influence. This outcome replicates previous reports that retrieval practice (as compared to restudy) reduces susceptibility to retroactive interference (Halamish & Bjork, 2011; Potts & Shanks, 2012; for related findings on retrieval-induced forgetting, see also Abel & Bäuml, 2014).

The main focus of Experiment 2 was again on the role of practice format, and the experimental conditions without retroactive interference serve as a conceptual replication of the pattern of results previously observed in the short-delay condition of Experiment 1. When a test was given after 5 min of an unrelated distractor task, retrieval practice (in comparison to restudy) was found to benefit recall, and importantly, this enhancement was not affected by whether the retrieval practice had been carried out in a blocked or a mixed fashion. Moreover, the data show that this finding extends to the conditions with retroactive interference. The results of Experiment 2 thus replicate the pattern from Experiment 1, in that the role of practice format was the same in the more difficult final-test conditions (long delay or retroactive interference) as in the immediate-test conditions.

General discussion

The present study reports two experiments that investigated the influence of blocked versus mixed practice on the mnemonic benefits of retrieval practice, applying procedures that are typical for testing-effect studies. In both experiments, recall was found to be enhanced after retrieval practice (plus feedback) in comparison to restudy, which is consistent with the testing-effect literature (see Roediger & Butler, 2011; Rowland, 2014). Most importantly, this retrieval-specific boost in memory performance was not affected by the exact format of practice (blocked vs. mixed)—neither in Experiment 1 (after a short or a long delay) nor in Experiment 2 (with or without retroactive interference). Together, the experiments indicate that the format of retrieval practice (blocked vs. mixed) does not affect its benefits, irrespective of whether they are assessed at a short-delay baseline, after prolonged retention intervals, or after the induction of retroactive interference.

The present experiments were primarily motivated by a prior study on retrieval-induced forgetting that had also investigated the role of blocked and mixed practice, and the results reported here may point to a difference between testing effects and retrieval-induced forgetting effects when they are investigated as a function of practice format. Of course, testing-effect studies are concerned with the beneficial effects of retrieval practice for the directly practiced contents, whereas studies on retrieval-induced forgetting usually focus on the detrimental influence of selective retrieval practice for related but unpracticed contents (see Anderson et al., 1994). However, randomly intermixing retrieval-practice and restudy trials seems to produce different outcomes on these positive and negative effects of retrieval practice. In particular, whereas retrieval-induced forgetting has repeatedly been shown to arise only after retrieval practice (but not after restudy) when the two practice types are blocked and clearly separable (e.g., Bäuml & Aslan, 2004; Ciranni & Shimamura, 1999; Wimber et al., 2009), a recent study by Dobler and Bäuml (2013) reported that random switches between retrieval-practice and restudy trials led to retrieval-induced forgetting after both types of practice. On the basis of prior task-switching work (see Allport et al., 1994; Campbell, 2005; Meuter & Allport, 1999), Dobler and Bäuml suggested that the task demands of the more difficult retrieval-practice trials might spill over to the easier restudy trials when practice is mixed. If so, subjects might engage in retrieval practice even during restudy trials.

The present results indicate that such dynamic effects between retrieval practice and restudy when practice is mixed are of far less importance for testing effects. In contrast to retrieval-induced forgetting, the testing effect seems to emerge exclusively after retrieval practice and does not spill over to restudy trials when practice is mixed. Thus, one possible interpretation of the present data could be to conclude that mixed practice does not increase retrieval in restudy conditions, contrary to Dobler and Bäuml's (2013) reasoning with regard to their retrieval-induced forgetting data. Alternatively, however, the different influences of mixed practice for the detrimental and beneficial effects of retrieval practice might emerge because random switches from retrieval-practice to restudy trials do not affect all to-be-restudied materials and/or all restudy trials. Because the benefits of retrieval practice may rely more critically on successful and repeated retrieval of the contents of memory (e.g., Butler & Roediger, 2007; Karpicke & Roediger, 2007), and because random switches might not sufficiently stimulate retrieval practice during the majority of restudy trials, the testing effect is not much affected by mixed practice. In contrast, for retrieval-induced forgetting, some previous studies have indicated that retrieval success and repeated retrieval may play a minor role only (Storm, Bjork, Bjork, & Nestojko, 2006; Storm & Nestojko, 2010; see also the results of the large-scale meta-analysis by Murayama et al., 2014). Because the two effects differ in this dimension, the testing effect may not be

affected by changes in the dynamics between retrieval practice and restudy, whereas retrieval-induced forgetting is.

A series of recent studies have examined the role of list composition for the testing effect under experimental conditions that were modeled to be similar to those of generation-effect studies. Although in one of these studies Rowland et al. (2014) reported that the magnitude of the testing effect was not affected by whether pure or mixed lists of retrieval-practice and restudy trials were used, a second study by Mulligan and Peterson (2015) that was also designed in parallel to studies on the generation effect reported a significant interaction, reflecting a larger testing effect with mixed than with blocked practice. The inconsistency in results across the two reports was addressed in another recent study conducted by Mulligan et al. (2016). There, four experiments confirmed that the testing effect can be affected by list composition in the same way as the generation effect, at least when experimental conditions are applied under which the generation effect is also expected to be larger for mixed than for pure lists. In particular, as was discussed by Mulligan et al., procedural details that are critical for whether or not effects of list composition will emerge (e.g., free recall as the final-test format and the application of separate free-recall tests after each list; see also the paragraph below) are often linked to the predictions of theoretical accounts of the examined effects (e.g., in this case, the item-specific–relational account or the item-order account of the generation effect; for details, see Mulligan et al., 2016; for a review, see McDaniel & Bugg, 2008). In sum, the results of the recent studies on the role of list composition indicate that there may indeed be certain parallels between the testing and generation effects, suggesting that theoretical accounts that have been discussed for the generation effect may also hold relevance for the testing effect and should be scrutinized in further studies (but see Karpicke & Zaromb, 2010).

In contrast to these previous studies, the procedures applied in the present experiments were purposefully chosen to be representative of typical testing-effect studies. Yet, consistent with the reasoning by Mulligan et al. (2016; see also the paragraph above), the present results confirmed that the testing effect is not larger after mixed than after blocked practice when the experimental conditions are not specifically modeled after those of generation-effect studies. In particular, generation-effect studies often use free recall as the criterial test and conduct separate free-recall tests after study of each single (pure or mixed) list, so that the retrieval sets are unique for each pure list and are not collapsed across conditions. In contrast, following typical studies on the testing effect, the present study applied one final cued-recall test for all of the practiced materials, conducted after a short delay, a prolonged delay, or in the presence of retroactive interference; under such more standard conditions, no evidence for larger testing effects with mixed practice arose.

Importantly, this difference in outcomes between the different sets of studies fits with the predictions of the item-specific–relational account or the item-order account of the generation effect (Hunt & McDaniel, 1993; McDaniel & Bugg, 2008; Mulligan & Lozito, 2004; Nairne et al., 1991). In essence, these accounts assume that, in mixed lists, more unusual or more demanding items (like to-be-generated or to-be-retrieved items) benefit from greater item-specific encoding, but at the same time they disrupt relational (i.e., order) encoding for all list items (including the more usual items that are read or restudied and that benefit from relational encoding). Due to this imbalance in encoding, the differences between unusual and usual items are much more pronounced or are exclusively present in mixed as compared to pure lists. A specific prediction of these accounts is that such sensitivity to list composition should be present with free-recall but not with cued-recall tests, because relational (order) information is not as important in cued recall. Past research confirmed this prediction for the generation effect (e.g., Burns, 1990, 1992). The present cued-recall results showing no list composition effect, together with Mulligan et al.'s (2016) free-recall results that showed intact list composition effects, indicate that the testing effect could follow the same prediction from the item-order account, as well. Of course, future work will be needed to provide more direct evidence in support of this idea.

For the present study we used a design typical for testing effects and found no evidence for list composition effects, but also no evidence for larger testing effects after blocked than after mixed practice, as would be expected on the basis of a prior study on retrieval-induced forgetting reported by Dobler and Bäuml (2013). Overall, the present results may indicate that the mnemonic benefits of retrieval practice as investigated in typical paired-associate testing-effect studies are robust across procedures in which retrieval practice is mixed with restudy trials or the two types of trials are blocked. This conclusion is also consistent with Rowland's (2014) recent meta-analysis on the testing effect, in which practice format was included as a moderator variable. In particular, in his analysis, 42 effect sizes obtained from testing-effect studies with mixed practice were compared to 87 effect sizes obtained from testing-effect studies applying blocked practice, and the results indicated that both practice formats resulted in reliable testing effects that did not differ in size. By directly manipulating practice format and obtaining similar results, the present study confirms this conclusion that was drawn on the basis of between-study contrasts. In addition, the fact that mixed or blocked practice format has no effect on the magnitude of the testing effect generalizes across whether the final test is comparatively easy (i.e., after short delay or in the absence of retroactive interference) or comparatively difficult (i.e., after a week's delay or in the presence of retroactive interference), which adds a new aspect to the literature.

To conclude, the present experiments show that the benefits of retrieval practice in typical testing-effect studies are not affected by whether the to-be-practiced material is presented in a blocked or a mixed fashion. This robustness of the testing effect to changes in the procedure may be especially important, given its potential to increase long-term retention in applied educational settings. In particular, researchers have recommended using retrieval practice both in the classroom and as a student study strategy to increase learning (e.g., Agarwal, Bain, & Chamberlain, 2012; Roediger, Putnam, & Smith, 2011). Because the testing effect does not seem to be affected by practice format under our conditions, the benefits are as great in blocked as in mixed practice. This should be comforting news, because our findings and others (e.g., Putnam & Roediger, 2013, or Smith, Roediger, & Karpicke, 2013, comparing response modes and the influences of covert vs. overt retrieval practice) have shown that the benefits of retrieval practice arise across many different variations of how exactly practice could be carried out. Retrieval practice boosts memory and can be applied as an effective study strategy by students; apart from stopping too early or practicing retrieval too little (e.g., Karpicke & Roediger, 2007; Vaughn, Rawson, & Pyc, 2013), there does not seem to be much that can be done wrong, even when retrieval practice is applied in various ways.

Author note This work was supported by a fellowship within the Postdoc Program of the German Academic Exchange Service (DAAD) and by a collaborative activity grant from the James S. McDonnell Foundation. The authors thank Andy DeSoto, Jason Finley, John Nestojko, Adam Putnam, and Victor Sungkhasettee for helpful discussions and comments on this project.

References

Abel, M., & Bäuml, K.-H. T. (2014). The roles of delay and retroactive interference in retrieval-induced forgetting. *Memory & Cognition, 42,* 141–150. doi:10.3758/s13421-013-0347-0

Agarwal, P. K., Bain, P. M., & Chamberlain, R. W. (2012). The value of applied research: Retrieval practice improves classroom learning and recommendations from a teacher, a principal, and a scientist. *Educational Psychology Review, 24,* 437–448.

Allport, A., Styles, E. A., & Hsieh, S. (1994). Shifting attentional set: Exploring the dynamic control of tasks. In C. Umiltà & M. Moscovitch (Eds.), *Attention and performance XV: Conscious and nonconscious information processing* (pp. 421–452). Cambridge, MA: MIT Press.

Anderson, M. C. (2003). Rethinking interference theory: Executive control and the mechanism of forgetting. *Journal of Memory and Language, 49,* 415–445. doi:10.1016/j.jml.2003.08.006

Anderson, M. C., Bjork, R. A., & Bjork, E. L. (1994). Remembering can cause forgetting: Retrieval dynamics in long-term memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 20,* 1063–1087. doi:10.1037/0278-7393.20.5.1063

Anderson, M. C., Bjork, E. L., & Bjork, R. A. (2000). Retrieval-induced forgetting: Evidence for a recall-specific mechanism. *Psychonomic Bulletin & Review, 7,* 522–530. doi:10.3758/BF03214366

Bäuml, K.-H., & Aslan, A. (2004). Part-list cuing as instructed retrieval inhibition. *Memory & Cognition, 32,* 610–617. doi:10.3758/BF03195852

Bäuml, K.-H. T., Holterman, C., & Abel, M. (2014). Sleep can reduce the testing effect—It enhances recall of restudied items but can leave recall of retrieved items unaffected. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 40,* 1568–1581.

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

Burns, D. J. (1990). The generation effect: A test between single- and multifactor theories. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 16,* 1060–1067. doi:10.1037/0278-7393.16.6.1060

Burns, D. J. (1992). The consequences of generation. *Journal of Memory and Language, 31,* 615–633.

Butler, A. C. (2010). Repeated testing produces superior transfer of learning relative to repeated studying. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 36,* 1118–1133. doi:10.1037/a0019902

Butler, A. C., & Roediger, H. L., III. (2007). Testing improves long-term retention in a simulated classroom setting. *European Journal of Cognitive Psychology, 19,* 514–527.

Campbell, J. I. D. (2005). Asymmetrical language switching costs in Chinese–English bilinguals' number naming and simple arithmetic. *Bilingualism: Language and Cognition, 8,* 85–91.

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

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

Carrier, M., & Pashler, H. (1992). The influence of retrieval on retention. *Memory & Cognition, 20,* 633–642. doi:10.3758/BF03202713

Cepeda, N. J., Pashler, H., Vul, E., Wixted, J. T., & Rohrer, D. (2006). Distributed practice in verbal recall tasks: A review and quantitative synthesis. *Psychological Bulletin, 132,* 354–380. doi:10.1037/0033-2909.132.3.354

Ciranni, M. A., & Shimamura, A. P. (1999). Retrieval-induced forgetting in episodic memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 25,* 1403–1414. doi:10.1037/0278-7393.25.6.1403

Cohen, J. (1988). *Statistical power analyses for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Congleton, A., & Rajaram, S. (2012). The origin of the interaction between learning method and delay in the testing effect: The roles of processing and conceptual retrieval organization. *Memory & Cognition, 40,* 528–539. doi:10.3758/s13421-011-0168-y

Dobler, I. M., & Bäuml, K.-H. T. (2013). Retrieval-induced forgetting: Dynamic effects between retrieval and restudy trials when practice is mixed. *Memory & Cognition, 41,* 547–557.

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

Gallistel, C. R. (2009). The importance of proving the null. *Psychological Review, 116,* 439–453. doi:10.1037/a0015251

Halamish, V., & Bjork, R. A. (2011). When does testing enhance retention? A distribution-based interpretation of retrieval as a memory modifier. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 37,* 801–812.

Hanslmayr, S., Staudigl, T., Aslan, A., & Bäuml, K.-H. (2010). Theta oscillations predict the detrimental effects of memory retrieval. *Cognitive, Affective, & Behavioral Neuroscience, 10,* 329–338. doi:10.3758/CABN.10.3.329

Hunt, R. R., & McDaniel, M. A. (1993). The enigma of organization and distinctiveness. *Journal of Memory and Language, 32,* 421–445. doi:10.1006/jmla.1993.1023

Jacoby, L. L., & Wahlheim, C. N. (2013). On the importance of looking back: The role of recursive remindings in recency judgments and cued recall. *Memory & Cognition, 41,* 625–637. doi:10.3758/s13421-013-0298-5

Karpicke, J. D., & Roediger, H. L., III. (2007). Repeated retrieval during learning is the key to long-term retention. *Journal of Memory and Language, 57,* 151–162. doi:10.1016/j.jml.2006.09.004

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

Karpicke, J. D., & Zaromb, F. M. (2010). Retrieval mode distinguishes the testing effect from the generation effect. *Journal of Memory and Language, 62,* 227–239.

Kornell, N., Bjork, R. A., & Garcia, M. A. (2011). Why tests appear to prevent forgetting: A distribution-based bifurcation model. *Journal of Memory and Language, 65,* 85–97.

Masson, M. E. J. (2011). A tutorial on a practical Bayesian alternative to null-hypothesis significance testing. *Behavior Research Methods, 43,* 679–690. doi:10.3758/s13428-010-0049-5

McDaniel, M. A., & Bugg, J. M. (2008). Instability in memory phenomena: A common puzzle and a unifying explanation. *Psychonomic Bulletin & Review, 15,* 237–255. doi:10.3758/PBR.15.2.237

McDermott, K. B., Arnold, K. M., & Nelson, S. M. (2014). The testing effect. In T. J. Perfect & D. S. Lindsay (Eds.), *The Sage handbook of applied memory* (pp. 183–200). Thousand Oaks, CA: Sage.

Meuter, R. F. I., & Allport, A. (1999). Bilingual language switching in naming: Asymmetrical costs of language selection. *Journal of Memory and Language, 40,* 25–40.

Mulligan, N. W., & Lozito, J. P. (2004). Self-generation and memory. In B. H. Ross (Ed.), *The psychology of learning and motivation: Advances in research and theory* (Vol. 45, pp. 175–214). San Diego, CA: Elsevier Academic Press.

Mulligan, N. W., & Peterson, D. J. (2015). Negative and positive testing effects in terms of item-specific and relational information. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 41,* 859–871. doi:10.1037/xlm0000056

Mulligan, N. W., Susser, J. A., & Smith, S. A. (2016). The testing effect is moderated by experimental design. *Journal of Memory and Language, 90,* 49–65.

Murayama, K., Miyatsu, T., Buchli, D., & Storm, B. C. (2014). Forgetting as a consequence of retrieval: A meta-analytic review of retrieval-induced forgetting. *Psychological Bulletin, 140,* 1383–1409.

Nairne, J. S., Riegler, G. L., & Serra, M. (1991). Dissociative effects of generation on item and order retention. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 17,* 702–709. doi:10.1037/0278-7393.17.4.702

Nelson, T. O., & Dunlosky, J. (1994). Norms of paired-associate recall during multitrial learning of Swahili–English translation equivalents. *Memory, 2,* 325–335. doi:10.1080/09658219408258951

Potts, R., & Shanks, D. R. (2012). Can testing immunize memories against interference? *Journal of Experimental Psychology: Learning, Memory, and Cognition, 38,* 1780–1785.

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

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

Raftery, A. E. (1995). Bayesian model selection in social research. In P. V. Marsden (Ed.), *Sociological methodology 1995* (pp. 111–196). Malden, MA: Blackwell.

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

Roediger, H. L., III, & Karpicke, J. D. (2006). Test-enhanced learning: Taking memory tests improves long-term retention. *Psychological Science, 17,* 249–255. doi:10.1111/j.1467-9280.2006.01693.x

Roediger, H. L., III, Putnam, A. L., & Smith, M. A. (2011). Ten benefits of testing and their applications to educational practice. In J. Mestre & B. Ross (Eds.), *The psychology of learning and motivation: Cognition in education* (Vol. 55, pp. 1–36). San Diego, CA: Elsevier Academic Press.

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

Rowland, C. A., Littrell-Baez, M. K., Sensenig, A. E., & DeLosh, E. L. (2014). Testing effects in mixed- versus pure-list designs. *Memory & Cognition, 42,* 912–921. doi:10.3758/s13421-014-0404-3

Slamecka, N. J., & Katsaiti, L. T. (1987). The generation effect as an artifact of selective displaced rehearsal. *Journal of Memory and Language, 26,* 589–607.

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

Storm, B. C., Bjork, E. L., Bjork, R. A., & Nestojko, J. F. (2006). Is retrieval success a necessary condition for retrieval-induced forgetting? *Psychonomic Bulletin & Review, 13,* 1023–1027. doi:10.3758/BF03194002

Storm, B. C., & Nestojko, J. F. (2010). Successful inhibition, unsuccessful retrieval: Manipulating time and success during retrieval practice. *Memory, 18,* 99–114. doi:10.1080/09658210903107853

Tullis, J. G., Benjamin, A. S., & Ross, B. H. (2014). The reminding effect: Presentation of associates enhances memory for related words in a list. *Journal of Experimental Psychology: General, 143,* 1526–1540.

Van Overschelde, J. P., Rawson, K. A., & Dunlosky, J. (2004). Category norms: An updated and expanded version of the Battig and Montague (1969) norms. *Journal of Memory and Language, 50,* 289–335. doi:10.1016/j.jml.2003.10.003

Vaughn, K. E., Rawson, K. A., & Pyc, M. A. (2013). Repeated retrieval practice and item difficulty: Does criterion learning eliminate item difficulty effects? *Psychonomic Bulletin & Review, 20,* 1239–1245. doi:10.3758/s13423-013-0434-z

Wagenmakers, E.-J. (2007). A practical solution to the pervasive problems of *p* values. *Psychonomic Bulletin & Review, 14,* 779–804. doi:10.3758/BF03194105

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

Wimber, M., Rutschmann, R. M., Greenlee, M. W., & Bäuml, K.-H. (2009). Retrieval from episodic memory: Neural mechanisms of interference resolution. *Journal of Cognitive Neuroscience, 21,* 538–549. doi:10.1162/jocn.2009.21043

Zaromb, F. M., & Roediger, H. L., III. (2010). The testing effect in free recall is associated with enhanced organizational processes. *Memory & Cognition, 38,* 995–1008. doi:10.3758/MC.38.8.995