Journal of Experimental Psychology:
Learning, Memory, and Cognition
2014, Vol. 40, No. 6, 1568–1581

0278-7393/14/$12.00 http://dx.doi.org/10.1037/xlm0000025

Sleep Can Reduce the Testing Effect: It Enhances Recall of Restudied Items but Can Leave Recall of Retrieved Items Unaffected

Karl-Heinz T. Bäuml, Christoph Holterman, and Magdalena Abel
Regensburg University

The testing effect refers to the finding that retrieval practice in comparison to restudy of previously encoded contents can improve memory performance and reduce time-dependent forgetting. Naturally, long retention intervals include both wake and sleep delay, which can influence memory contents differently. In fact, sleep immediately after encoding can induce a mnemonic benefit, stabilizing and strengthening the encoded contents. We investigated in a series of 5 experiments whether sleep influences the testing effect. After initial study of categorized item material (Experiments 1, 2, and 4A), paired associates (Experiment 3), or educational text material (Experiment 4B), subjects were asked to restudy encoded contents or engage in active retrieval practice. A final recall test was conducted after a 12-hr delay that included diurnal wakefulness or nocturnal sleep. The results consistently showed typical testing effects after the wake delay. However, these testing effects were reduced or even eliminated after sleep, because sleep benefited recall of restudied items but left recall of retrieved items unaffected. The findings are consistent with the bifurcation model of the testing effect (Kornell, Bjork, & Garcia, 2011), according to which the distribution of memory strengths across items is shifted differentially by retrieving and restudying, with retrieval strengthening items to a much higher degree than restudy does. On the basis of this model, most of the retrieved items already fall above recall threshold in the absence of sleep, so additional sleep-induced strengthening may not improve recall of retrieved items any further.

Keywords: testing effect, retrieval, restudy, sleep

The act of retrieving information from memory is a powerful tool to promote long-term retention, as is demonstrated in studies on the testing effect. In testing effect studies, subjects are usually asked to either actively retrieve or passively restudy previously presented contents. The typical finding is that retrieval practice in comparison to restudy improves memory performance and reduces time-dependent forgetting (Hogan & Kintsch, 1971; Roediger & Karpicke, 2006; for a review, see Roediger & Butler, 2011). The testing effect has been observed with a variety of materials, like word lists (Hogan & Kintsch, 1971), paired associate lists (Carrier & Pashler, 1992), picture lists (Wheeler & Roediger, 1992), and educational text material (Roediger & Karpicke, 2006). It can be modulated by a number of factors, like retention interval between retrieval and test (Roediger & Karpicke, 2006; Toppino & Cohen, 2009), difficulty of retrieval task (Carpenter & DeLosh, 2006; Kornell, Bjork, & Garcia, 2011), and final test format (Halamish & Bjork, 2011; Hogan & Kintsch, 1971).

To date, several theoretical accounts have been proposed to explain why retrieval promotes long-term retention (see Roediger & Butler, 2011). One influential account is the elaborative retrieval hypothesis (Carpenter, 2009; McDaniel & Masson, 1985; Pyc & Rawson, 2010), according to which retrieval practice induces more elaborative processing than restudy does. For instance, when one is attempting to retrieve a target item from memory, semantically related items may be activated during the search for the target information and become linked to the target item (Carpenter, 2009; Pyc & Rawson, 2010, 2012). Such extra information may be activated mainly during more difficult retrieval tasks, when the target information is less readily retrievable and more extensive memory search is required, and may be less activated or not activated at all during easier retrieval tasks or restudy opportunities, when the target information is more easily retrieved or is even reexposed intact. Consistent with this view, the testing effect has sometimes been found to be larger after more difficult versus easier retrieval tasks (Carpenter, 2009; Kornell et al., 2011). Because the encoding of the extra semantic information may become beneficial mainly after a prolonged retention interval, when retrieval supposedly becomes more semantic in nature (Carpenter, 2011), the account can also explain why the size of the testing effect typically increases with retention interval.

Another, more recent account of the testing effect is the bifurcation model (Halamish & Bjork, 2011; Kornell et al., 2011). In contrast to the elaborative retrieval hypothesis, this model is not intended to be a process model that provides an explanation of the mechanisms underlying the testing effect. Rather, its goal is to account for why certain factors modulate the testing effect, given that successful testing promotes retention more than restudy does.

This article was published Online First June 16, 2014.

Karl-Heinz T. Bäuml, Christoph Holterman, and Magdalena Abel, Department of Experimental Psychology, Regensburg University.

This work was supported by Grant BA 1382/14-1 from the German Research Foundation (Deutsche Forschungsgemeinschaft) awarded to Karl-Heinz T. Bäuml. We thank R. A. Bjork, V. Halamish, H. L. Roediger III, and M. K. Scullin for their comments on a previous version of the article. We also thank V. Haller, A. Karl, C. Nottberg, A. Schlichting, D. Schnell, and F. Welker for their help with data collection.

Correspondence concerning this article should be addressed to Karl-Heinz T. Bäuml, Department of Experimental Psychology, Regensburg University, 93040 Regensburg, Germany. E-mail: karl-heinz.baeuml@ur.de

At the core of the model is the assumption that, on a scale of memory strength, retrieval creates a bifurcated distribution of items, with the successfully retrieved items being strengthened to a very high degree and the nonretrieved items remaining at their original strength level. In contrast, restudy is assumed to strengthen all restudied items about equally, although to a lower degree than successful retrieval does for retrieved items (see Figure 1). Although the bifurcation model assumes that retrieval-practiced and restudied items decrease in strength with increasing delay at a comparable rate, it can explain the testing effect. Because of their suggested high strength level, many of the successfully retrieved items may remain above recall threshold after both short and prolonged retention intervals; in contrast, because of their suggested lower strength level, many of the restudied items may remain above recall threshold after short but no longer after prolonged retention interval. Assuming that difficult test formats lead to higher recall thresholds than easy test formats do, the model can also explain why the testing effect has been found to be larger in the presence of retroactive interference (difficult test format) than in its absence (easy test format; Halamish & Bjork, 2011). An increasing threshold should leave recall chances for many of the highly strengthened retrieved items largely unaffected but should reduce recall chances for many of the less well strengthened restudied items (for details, see Halamish & Bjork, 2011).[1]

Not only retrieval but also sleep that directly follows encoding can reduce time-dependent forgetting (e.g., Barrett & Ekstrand, 1972; Gais, Lucas, & Born, 2006). Using a variety of study materials, like word lists (Ficca, Lombardo, Rossi, & Salzarulo, 2000), paired associates (Plihal & Born, 1997), or spatial information (Talamini, Nieuwenhuis, Takashima, & Jensen, 2008), numerous studies have indeed documented the beneficial effect of sleep over wake delay on memory performance. Although early theories assumed that sleep was a rather passive state, sheltering memories from interference that would otherwise accumulate during wakefulness (Jenkins & Dallenbach, 1924), results from newer studies report evidence that memory contents are reactivated during sleep. Such reactivation seems to foster sleep-associated memory consolidation by actively stabilizing and strengthening the memory contents (for reviews, see Diekelmann & Born, 2010; Stickgold & Walker, 2013).

The testing effect has typically been reported for delay intervals of several days, with the retention interval naturally including both wake and sleep delay (e.g., Kornell et al., 2011; Pyc & Rawson, 2012; Roediger & Karpicke, 2006). Whether sleep that directly follows encoding affects retrieved and restudied memories differently and thus influences the testing effect has not been investigated to date. There are at least two empirical reasons, however, why sleep may indeed affect the testing effect. The more general reason is that sleep compared with being awake does not benefit all memories equally, but it can induce selective benefits for certain memories (see Stickgold & Walker, 2013). Consistently, emotional memories have been found to show more sleep benefits than neutral memories do (Payne, Stickgold, Swanberg, & Kensinger, 2008), and memories considered relevant for the future, compared with supposedly irrelevant material, show more sleep benefits (Wilhelm et al., 2011). For instance, if testing reduced the expectancy of a future memory test (Szpunar, McDermott, & Roediger, 2007), then sleep might be more beneficial for restudied than retrieved items and reduce the testing effect. The other, more specific reason is that a recent study provided evidence that retrieved items may not show sleep benefits, whereas nonretrieved control items may do so (Abel & Bäuml, 2012; for a similar finding, see Racsmány, Conway, & Demeter, 2010; see also Tucker & Fishbein, 2008). The finding raises the possibility that sleep may improve recall of restudied but not of retrieved items, which would reduce the testing effect. Because the result was based on a single experiment that included a study-only but no restudy condition as a control, however, caution is warranted before drawing more firm conclusions on the role of sleep for the testing effect.

Knowledge about the role of sleep for the testing effect may not only be of practical relevance but may also be of relevance for theories of the effect. For instance, although the elaborative retrieval hypothesis in itself makes no direct predictions regarding the influence of sleep and wake delay on the testing effect, results from several previous studies suggest that on the basis of this hypothesis, sleep may be expected to maintain or even increase the effect. According to this hypothesis, the testing effect benefits from the activation of extra semantic information during retrieval practice, but memory for semantic information has been found to be enhanced by sleep. For instance, while investigating the influence of sleep on false memories in the Deese–Roediger–McDermott paradigm (Roediger & McDermott, 1995), researchers conducting several studies found sleep facilitated not only veridical recall of studied noncritical items (e.g., *sugar, bitter, taste*) but recall of the semantically related but unstudied critical item (e.g., *sweet*), as well; moreover, although some of these studies reported similar beneficial effects of sleep for studied noncritical and unstudied critical items (e.g., Darsaud et al., 2011), others found even larger sleep benefits for the critical items (McKeon, Pace-Schott, & Spencer, 2012; Payne et al., 2009; but see Fenn, Gallo, Margoliash, Roediger, & Nusbaum, 2009). There is further evidence that sleep primes semantic associative networks when applying the remote associates test, showing that sleep can increase creativity compared with wakefulness (Cai, Mednick, Harrison, Kanady, & Mednick, 2009).

In contrast, on the basis of the bifurcation model, sleep may be expected to reduce the testing effect. Because the model assumes that successfully retrieved items (but not the nonretrieved items) are strengthened to a very high degree and restudied items to a relatively lower degree, mainly recall of restudied items should benefit from sleep. Indeed, if sleep increases the strength of restudied items, then after prolonged retention interval, a larger proportion of restudied items should fall above recall threshold after sleep than after waking. In contrast, because most of the successfully retrieved items already fall above recall threshold in the absence of sleep, additional sleep-induced strengthening of the

[1] Research on retrieval-induced forgetting demonstrates that retrieval of some studied items can impair recall of other items (Anderson, Bjork, & Bjork, 1994). Because the effect may be due to some weakening of the nonretrieved items' memory strength (Anderson, 2003; Bäuml, Pastötter, & Hanslmayr, 2010; but see Raaijmakers & Jakab, 2013), the assumption of the bifurcation model that nonretrieved items remain at the original strength level may not hold in general, and the distribution of nonretrieved items may need some shifting to the left (see also Halamish & Bjork, 2011). However, because the focus of the present study is mainly on restudied and retrieved items and less on the nonretrieved items, the issue is not of much relevance for the present work.

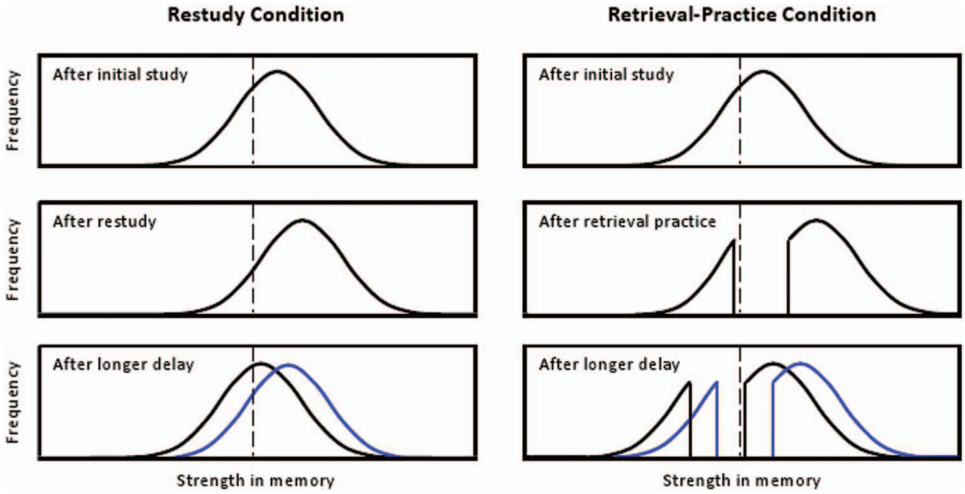

Figure 1. Illustration of memory strength distributions of two hypothetical sets of items, following Kornell, Bjork, and Garcia's (2011) bifurcation model. The left column shows items that were restudied and the right column items that were retrieval practiced. In the top pair of panels, memory strength after one initial study trial is illustrated; at this point, the two distributions are identical. The second pair of panels shows how distributions are shifted after restudy or retrieval practice. Although all restudied items gain memory strength, retrieval-practiced items become bifurcated. Successfully retrieved items are strengthened to a higher degree than restudied items are, whereas nonretrieved items remain at the original strength level. The bottom pair of panels illustrates how distributions may be shifted after wake delay (black curves) and sleep delay (blue [gray] curves), respectively; all items show the same amount of delay-induced forgetting and the same amount of sleep-induced strengthening. Vertical dotted lines indicate recall threshold at test. Restudied items profit from sleep, because more of them cross threshold after sleep. In contrast, retrieved items do not profit from sleep, because their strength level is still above threshold after wake delay. See the online article for the color version of this figure.

items should have only a small, if any, beneficial effect on recall performance, thus reducing the testing effect after sleep (see Figure 1).[2]

Experiment 1

Our goal in the present study was to examine whether sleep influences the testing effect by investigating the effects of sleep and wake delay on items that immediately after study were subject to retrieval-practice or restudy cycles. In Experiment 1, subjects initially studied a semantically categorized item list. In the restudy condition, subjects were then asked to restudy the items of half of the categories once and to restudy the items of the other half twice; analogously, subjects in the retrieval-practice condition were asked to retrieve the items of one half of the categories once and to retrieve the items of the other half twice. A final memory test was conducted after a 12-hr delay that included either regular sleep or wakefulness. With regard to the wake delay, we expected the typical testing effect finding, that is, better recall after retrieval practice than after restudy, both after one and after two practice cycles. With regard to the sleep delay, expectations depend on testing effect theory. On the basis of the elaborative retrieval hypothesis, sleep may be expected to maintain or even increase the testing effect. On the basis of the bifurcation model, sleep may be expected to reduce the testing effect.

Method

Participants. Originally, 224 students from Regensburg University were recruited for the experiment. Eight participants had to be excluded from the sample prior to data analysis because they reported either alcohol intake or daytime napping between sessions. A final sample of 216 healthy participants remained (M = 22.4 years; range: 18–30 years; 55 male). All participants of the final sample were native German speakers and were distributed equally across conditions (n = 36 in each of the six conditions). Comparisons between conditions regarding subjects' age, habitual sleep duration, subjective ratings of sleep quality, IQ (estimated via speed of cognitive processing; Oswald & Roth, 1987), and ratings on the Epworth Sleepiness Scale (Johns, 1991) did not reveal any differences (all ps $> .10$). Corresponding comparisons were calculated for all experiments presented in this article. Because they never revealed any significant differences between conditions, they will not be reported for each single experiment.

Material. An item list was constructed consisting of 24 concrete German nouns from four different semantic categories (six items per category; Scheithe & Bäuml, 1995; Van Overschelde, Rawson, & Dunlosky, 2004). Within each category, items had unique initial letters. Two of the categories were repeated once

[2] Although in Figure 1, the effect of sleep on restudied and retrieved items is shown by shifting the distributions of the two item types to the same extent, in principle, the bifurcation model can be combined with any assumption about the relative sleep-induced strengthening of restudied and retrieved items. It is important to note, however, that the arguments in the present study do not depend much on the exact shifting, because, on the basis of the bifurcation model, sleep should reduce the testing effect regardless of the amount of sleep-induced strengthening for the retrieved items, at least as long as sleep strengthens the restudied items (see also the General Discussion section).

during retrieval practice or restudy, whereas the remaining categories were repeated twice; categories were distributed across practice levels in a balanced manner.

Design. The experiment had a 2 × 2 × 2 mixed-factorial design with the between-subjects factors of type of practice (restudy, retrieval practice) and delay (12-hr wake, 12-hr sleep) and the within-subjects factor of practice level (low, high). After initial study, half of the subjects were asked to restudy the list (restudy condition), whereas the other half was asked to engage in retrieval practice (retrieval-practice condition). For half of the initially studied categories, subjects were given one restudy or retrieval-practice cycle (low practice level); the other half were given two restudy or retrieval-practice cycles (high practice level). In the 12-hr wake condition, participants studied and practiced the items at 9 a.m. and final test was conducted at 9 p.m., after 12 hr of wakefulness; in contrast, in the 12-hr sleep condition, participants studied and practiced the items at 9 p.m. and took the final test at 9 a.m., after one night of nocturnal sleep (see Figure 2; for similar designs, see Abel & Bäuml, 2013; Payne et al., 2008; Scullin & McDaniel, 2010). Because learning and test sessions took place at different times of day across delay conditions, an additional short-delay condition was included to control for potential circadian effects. Half of the subjects in this condition participated at 9 a.m., the other half at 9 p.m., with only a short delay of 12 min between learning phase and test.

Procedure.

Study and practice phase. During study, items were presented successively and together with their corresponding category labels in a random order, at a presentation rate of 3 s per item. After one initial study cycle for all items, additional practice cycles made up either restudy or retrieval-practice trials, depending on practice condition. In the restudy condition, the intact items and their category labels were reexposed at a 3-s rate. Items of two of the four categories were restudied once (SS = study-study), whereas items from the other two categories were restudied twice (SSS = study-study-study). In the retrieval condition, subjects were provided with the items' category labels and word stems and were asked to recall the corresponding items. When a response was made, the item was removed and the next item was presented. If participants did not respond within 5 s, the item was removed without a response being entered, and the next item was presented. Mean response time (which equals overall processing time) across all retrieval-practice trials was 2.1 s (SD = 0.27). Retrieval of two categories' items was practiced once (ST = study-test), whereas retrieval of the other two categories' items was practiced twice (STT = study-test-test). In both the restudy and the retrieval conditions, order of items was random with the restriction that items of the same category were never restudied or retrieved consecutively.

The learning phase was followed by a distractor phase of 12 min, during which participants engaged in several unrelated cognitive tasks. Afterward, subjects in the short-delay control conditions completed the final recall test. In contrast, subjects in the 12-hr delay conditions were dismissed from the first session after having completed half of the distractor phase (6 min). After a delay of 12 hr that was either spent awake or filled with normal nighttime sleep, subjects in the 12-hr delay conditions returned to the laboratory and completed the second half of the distractor phase (6 min) before completing the same final recall test. Concerning compliance with instructions, subjects in the 12-hr sleep conditions reported to have slept regularly during the night (M = 7.9 hr, SD = 0.83), whereas subjects in the 12-hr wake conditions reported that they had not taken naps during the day. None of the subjects reported alcohol intake between sessions.

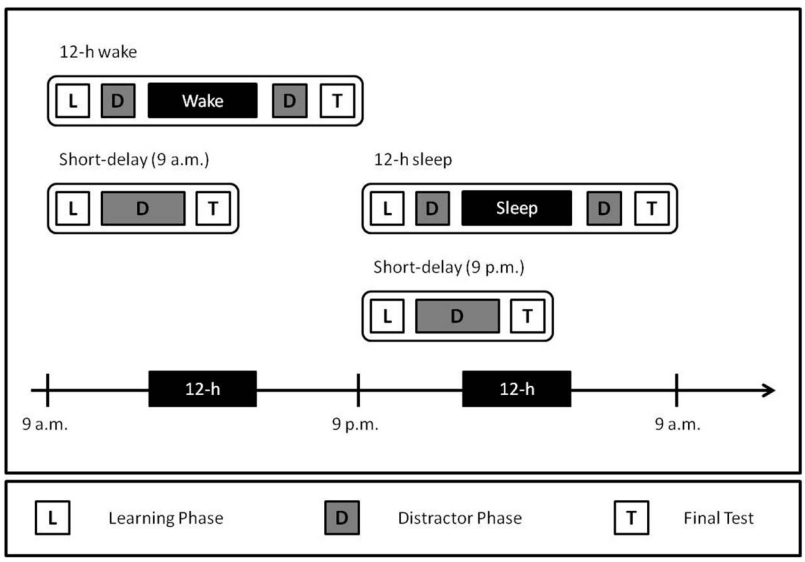

Figure 2. Illustration of conditions in Experiment 1: In the 12-hr wake condition, the learning of the material took place at 9 a.m., before participants returned to the laboratory for the final test after a 12-hr delay comprising daytime wakefulness. In the 12-hr sleep condition, the learning of the material took place at 9 p.m., and memory was tested after a 12-hr delay comprising nighttime sleep. In the short-delay condition, half of the subjects participated at 9 a.m.; the other half did so at 9 p.m. In this condition, the final test was administered after a short delay of 12 min.

Test phase. At test, participants were presented with the category labels and initial letters of all 24 studied items for 7 s each and were asked to recall the appropriate item. Items from the same category were tested consecutively. Order of categories and order of items within categories was random.

Results

Success rates during retrieval-practice cycles. A 2 × 2 × 2 analysis of variance (ANOVA) with the factors of time of day (9 a.m., 9 p.m.), retrieval practice (ST, STT), and delay (short delay, 12-hr delay) showed that retrieval success was higher after two than after one retrieval-practice cycle (95.3% vs. 89.8%), $F(1, 104) = 27.36$, $MSE = 56.61$, $p < .001$, $\eta^2 = .21$. There were no other main effects and no interactions, $ps > .10$.

Final test (12-hr delay conditions). Figure 3 shows recall performance after the 12-hr delay. A 2 × 2 × 2 ANOVA with the factors of type of practice (restudy, retrieval practice), delay (12-hr wake, 12-hr sleep), and practice level (low, high) revealed significant main effects of type of practice, $F(1, 140) = 12.13$, mean square error (MSE) = 274.67, $p = .001$, $\eta^2 = .08$; delay, $F(1, 140) = 19.22$, $MSE = 274.67$, $p < .001$, $\eta^2 = .12$; and practice level, $F(1, 140) = 34.79$, $MSE = 128.85$, $p < .001$, $\eta^2 = .20$. The main effect of type of practice reflects overall higher recall in the retrieval-practice condition than in the restudy condition (74.7% vs. 67.9%), whereas the main effect of delay shows that recall rates were higher overall in the 12-hr sleep condition than in the 12-hr wake condition (75.6% vs. 67.0%); the main effect of practice level indicates better recall after the high than after the low practice level (75.2% vs. 67.3%). More important, although all other interactions were nonsignificant (all $ps > .05$), there was a reliable two-way interaction between type of practice and delay, $F(1, 140) = 7.75$, $MSE = 275.67$, $p = .006$, $\eta^2 = .05$, indicating that the difference in memory performance after retrieval practice in comparison to restudy was modulated by delay condition. Consistently, planned comparisons showed a significant testing effect after the 12-hr wake delay, both for the lower practice level (57.9% vs. 67.6%), $t(70) = 2.53$, $p = .013$, $d = 0.60$, and the higher practice level (63.9% vs. 78.7%), $t(70) = 4.44$, $p < .001$, $d = 1.05$, whereas no reliable testing effect arose after the 12-hr sleep delay, for both the lower practice level (70.8% vs. 73.1%), $t(70) = 0.67$, $p = .505$, $d = 0.16$, and the higher practice level (78.9% vs. 79.4%), $t(70) = 0.87$, $p = .872$, $d = 0.04$. Further planned comparisons revealed that there was a beneficial effect of sleep for items that had been restudied, irrespective of whether they were restudied once (SS), $t(70) = 3.66$, $p < .001$, $d = 0.86$, or twice (SSS), $t(70) = 4.35$, $p < .001$, $d = 1.03$, whereas there was no such effect for items that had been subject to retrieval practice once (ST), $t(70) = 1.50$, $p = .138$, $d = 0.35$, or twice (STT), $t(70) = 0.27$, $p = .789$, $d = 0.06$.

Circadian control (short-delay condition). Table 1 shows mean recall levels after the short delay. A 2 × 2 × 2 ANOVA with the factors of type of practice (restudy, retrieval practice), time of day (9 a.m., 9 p.m.), and practice level (low, high) revealed a significant main effect of practice level, $F(1, 68) = 8.67$, $MSE = 121.26$, $p = .004$, $\eta^2 = .11$, with the items of the higher practice level being recalled better than the items of the lower practice level (81.4% vs. 76.0%). No other effects emerged, indicating that there was no testing effect after short retention interval and recall was unaffected by circadian effects, all $ps > .10$.

Discussion

The results of Experiment 1 replicate prior testing effect studies. Recall after the 12-hr wake delay was better after retrieval practice than after restudy, both after one and after two practice trials. In contrast, after the 12-hr sleep delay, the testing effect was reduced or even eliminated, again both after one and after two practice trials. The observed reduction of the testing effect arose because there was a beneficial effect of sleep after restudy trials but no such effect after retrieval practice trials. Empirically, these findings are consistent with recent results indicating no beneficial effect of sleep after retrieval practice (Abel & Bäuml, 2012). Theoretically, the findings agree with the bifurcation model. This model suggests beneficial effects of sleep on recall mainly after restudy trials and less after retrieval trials, which is what the present results show. The findings appear less consistent with the elaborative retrieval hypothesis, according to which beneficial effects of sleep on recall may be expected to be at least as large after retrieval as after restudy cycles.

Experiment 2

To the best of our knowledge, Experiment 1 is the first demonstration that sleep can reduce or even eliminate the testing effect. However, it could be argued that the elimination of the testing

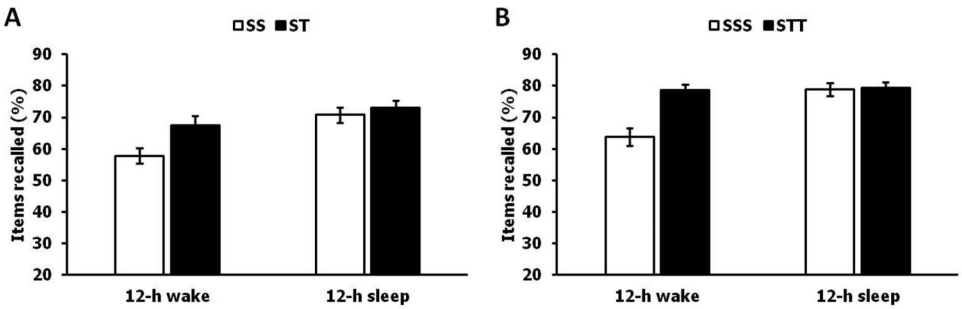

Figure 3. Mean recall performance on the final test of Experiment 1 as a function of delay (12-hr wake, 12-hr sleep) and type of practice (restudy, retrieval practice), separately for the low (A) and high (B) practice levels. Condition labels indicate study (S) and retrieval-practice (T) cycles. Error bars represent standard errors.

Table 1
Mean Recall Performance in the Short-Delay Condition of Experiment 1 as a Function of Time of Day, Type of Practice, and Practice Level

	Restudy				Retrieval practice			
	SS		SSS		ST		STT	
Time of day	*M*	*SE*	*M*	*SE*	*M*	*SE*	*M*	*SE*
9 a.m.	75.4	3.5	79.9	3.4	77.7	2.7	82.8	3.3
9 p.m.	73.9	3.8	78.1	2.7	77.3	4.0	83.3	2.7
Combined	74.5	2.5	79.8	2.2	77.5	2.4	83.1	2.0

Note. Condition labels indicate study (S) and retrieval-practice (T) cycles.

effect in Experiment 1 was due to a ceiling effect in the retrieval-practice condition, because mean recall levels were relatively high. If so, the effect might be observable after sleep if recall at test was more difficult than it was in Experiment 1. Halamish and Bjork (2011) increased difficulty at test by inducing retroactive interference and found the testing effect to be larger in the presence than in the absence of retroactive interference (for related results, see Abel & Bäuml, 2014; Potts & Shanks, 2012). Such a pattern is consistent with the bifurcation model, which assumes that recall threshold increases with test difficulty (see Halamish & Bjork, 2011).

Experiment 2 is similar to Experiment 1, with the main exception that retroactive interference was induced at test. After the 12-hr delay interval, an additional item list, consisting of items from the same categories as were used during original study, was presented for study. Study of the list should induce retroactive interference, thus making the recall task more difficult and preventing any ceiling effects on mean recall levels. Again, we expected to find the testing effect after the wake delay, the size of which, following Halamish and Bjork (2011), might be increased relative to Experiment 1. More important, if results in the retrieval condition of Experiment 1 were due to a ceiling effect for mean recall levels, then the testing effect in this experiment might arise after both wake and sleep delays, with similar size in the two delay conditions or even increased size after sleep compared with wake delays. Alternatively, if the results of Experiment 1 were not due to such a ceiling effect, then the testing effect may again be reduced or even eliminated after sleep.

Method

Participants. The final sample consisted of 144 healthy students from Regensburg University ($M = 22.9$ years; range: 18–30 years; 44 male; five additionally tested participants had to be excluded from the sample because of alcohol consumption or daytime napping between sessions). Again, all participants were native German speakers and were distributed equally across conditions ($n = 36$ in each of the four conditions).

Material. Two item lists were constructed: a target list and a nontarget list. The target list consisted of the same 24 nouns that were already used in Experiment 1. Additionally, to induce retroactive interference, we constructed a nontarget list with another 24 concrete German nouns, which were taken from the same four semantic categories that were included in the target list (six items per category; Scheithe & Bäuml, 1995; Van Overschelde et al., 2004). Within each category, target and nontarget items had unique initial letters.

Design. The experiment had a 2 × 2 × 2 mixed-factorial design with the between-subjects factors of type of practice (restudy, retrieval practice) and delay (12-hr wake, 12-hr sleep) and the within-subjects factor of practice level (low, high). Experiment 2 was identical to Experiment 1 with two exceptions: (a) the nontarget list was studied at the beginning of the second session and (b) no short-delay condition was included.

Procedure. Study and practice phases were conducted in exactly the same way as in Experiment 1. Participants were asked to study the target list on one initial study trial, which was followed by one or two restudy cycles (restudy condition) or one or two retrieval-practice cycles (retrieval-practice condition). In the retrieval condition, subjects were again provided with the items' category labels and word stems for up to 5 s each and were asked to recall the corresponding items. This time, mean response time (equaling overall processing time) across all retrieval-practice trials was 2.2 s ($SD = 0.44$). After the 12-hr delay, instead of using the full second half of the distractor phase, we presented the nontarget list. The list was presented three times at a rate of 3 s per item; items were presented in a random order. This phase was followed by a 30-s backward-counting distractor task. Afterward, the final test was conducted in a manner analogous to Experiment 1. The items of the target list were tested first and the items of the nontarget list second.

Concerning compliance with instructions, subjects in the 12-hr sleep conditions again reported to have slept regularly during the night ($M = 7.8$ hr, $SD = 0.93$), whereas subjects in the 12-hr wake conditions reported not to have taken naps during the day. None of the subjects reported alcohol intake between sessions.

Results

Success rates during retrieval-practice cycles. A 2 × 2 ANOVA with the factors of time of day (9 a.m., 9 p.m.) and retrieval practice (ST, STT) showed that retrieval success was higher after two than after one retrieval-practice cycle (96.5% vs. 90.0%), $F(1, 70) = 24.21$, $MSE = 62.83$, $p < .001$, $\eta^2 = .26$. There was no main effect of time of day and no interaction, $ps > .10$.

Final test (12-hr delay conditions). Figure 4 shows recall performance after the 12-hr delay. A 2 × 2 × 2 ANOVA with the factors of type of practice (restudy, retrieval practice), delay (12-hr wake, 12-hr sleep), and practice level (low, high) revealed significant main effects of type of practice, $F(1, 140) = 53.00$, $MSE = 410.44$, $p < .001$, $\eta^2 = .28$; delay, $F(1, 140) = 9.26$, $MSE = 410.44$, $p = .003$, $\eta^2 = .06$; and practice level, $F(1, 140) = 33.60$, $MSE = 171.53$, $p < .001$, $\eta^2 = .19$. The main effect of type of practice again reflects overall higher recall in the retrieval condition than in the restudy condition (68.4% vs. 51.0%), whereas the main effect of delay shows that recall rates were overall higher in the 12-hr sleep condition than in the 12-hr wake condition (63.4% vs. 56.1%); the main effect of practice level indicates that items were recalled better with the higher than the lower practice level (64.2% vs. 55.3%). More important, although all other interactions were nonsignificant (all $ps > .05$), there was once more a reliable two-way interaction between type of practice and delay, $F(1,$

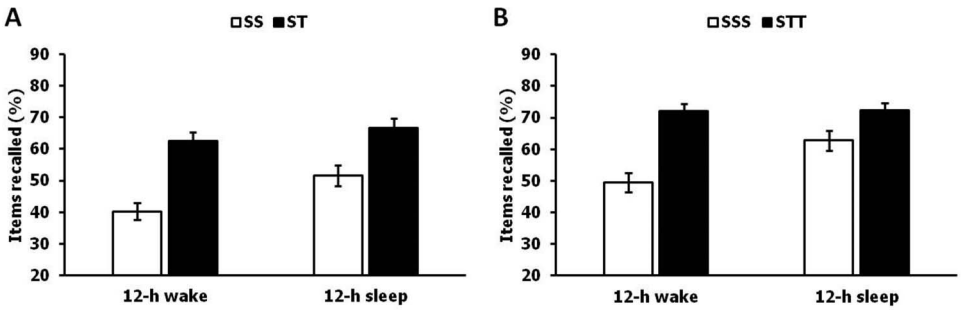

Figure 4. Mean recall performance of target items on the final test of Experiment 2 as a function of delay (12-hr wake, 12-hr sleep) and type of practice (restudy, retrieval practice), separately for the low (A) and high (B) practice levels. Condition labels indicate study (S) and retrieval practice (T) cycles. Error bars represent standard errors.

140) = 4.39, *MSE* = 410.44, $p = .038$, $\eta^2 = .03$, indicating that the beneficial effect of retrieval practice in comparison to restudy was again modulated by delay condition. Consistently, planned comparisons revealed significant testing effects after the 12-hr wake delay for both the lower practice level (40.3% vs. 62.4%), $t(70) = 5.57$, $p < .001$, $d = 1.31$, and the higher practice level (49.5% vs. 72.1%), $t(70) = 6.01$, $p < .001$, $d = 1.42$, but numerically reduced testing effects after the 12-hr sleep delay, for both the lower practice level (51.6% vs. 66.7%), $t(70) = 3.49$, $p = .001$, $d = 0.82$, and the higher practice level (62.7% vs. 72.4%), $t(70) = 2.43$, $p = .018$, $d = 0.57$. More detailed analysis further showed that there was a beneficial effect of sleep for items that had been restudied once (SS), $t(70) = 2.73$, $p = .008$, $d = 0.64$, or twice (SSS), $t(70) = 3.03$, $p = .003$, $d = 0.71$, whereas there was no such effect for items that had been retrieval practiced once (ST), $t(70) = 1.02$, $p = .311$, $d = 0.24$, or twice (STT), $t(70) = 0.08$, $p = .934$, $d = 0.02$.

Recall for the nontarget list was analyzed in a 2 × 2 ANOVA with the factors of type of practice (restudy, retrieval practice) and delay (12-hr wake, 12-hr sleep). The analysis revealed no significant effects, all $Fs < 1.0$, indicating that there were no reliable differences between practice conditions (58.6% vs. 57.9%) and no reliable differences between delay conditions (58.2% vs. 58.2%).

Additional analyses. The main difference between Experiments 1 and 2 is that retroactive interference was induced in Experiment 2 but not in Experiment 1. We therefore directly compared results of the two experiments by means of a 2 × 2 × 2 × 2 ANOVA with the factors of type of practice (restudy, retrieval practice), delay (12-hr wake, 12-hr sleep), practice level (low, high), and interference (interference, no interference). This analysis again revealed significant main effects of type of practice, $F(1, 280) = 61.47$, *MSE* = 342.55, $p < .001$, $\eta^2 = .18$; delay, $F(1, 280) = 26.33$, *MSE* = 342.55, $p < .001$, $\eta^2 = .09$; and practice level, $F(1, 280) = 67.95$, *MSE* = 150.19, $p < .001$, $\eta^2 = .20$, as well as a significant interaction between type of practice and delay, $F(1, 280) = 11.46$, *MSE* = 342.55, $p = .001$, $\eta^2 = .04$. In addition, it revealed a significant main effect of interference, $F(1, 280) = 56.11$, *MSE* = 352.55, $p < .001$, $\eta^2 = .17$, indicating lower recall rates in Experiment 2 than in Experiment 1 (59.7% vs. 71.3%). Most important, there was a significant interaction between type of practice and interference, $F(1, 280) = 11.76$, *MSE* = 342.55, $p = .001$, $\eta^2 = .04$, indicating that the testing effect was influenced by interference. Indeed, planned comparisons revealed that the testing effect for the lower practice level was greater in the presence of interference (45.9% vs. 64.6%), $t(142) = 6.20$, $p < .001$, $d = 1.03$, than in its absence (64.3% vs. 70.3%), $t(142) = 2.24$, $p = .027$, $d = 0.37$, and, similarly, the testing effect for the higher practice level was greater in the presence of interference (56.1% vs. 72.3%), $t(142) = 5.70$, $p < .001$, $d = 0.95$, than in its absence (71.4% vs. 79.0%), $t(142) = 3.28$, $p = .001$, $d = 0.55$. No further effects were significant ($ps > .10$).

Discussion

By inducing retroactive interference at test, Experiment 2 used a more difficult recall test than Experiment 1 did. Doing so, it successfully prevented ceiling effects on mean recall levels in the retrieval practice condition and replicated prior work by showing larger testing effects in the presence than in the absence of retroactive interference (Halamish & Bjork, 2011). More important, consistent with the results of Experiment 1, the results of Experiment 2 showed reduced testing effects after the sleep delay compared with the wake delay, although here reliable effects arose after both wake and sleep delays. Like in Experiment 1, the reduction of the testing effect after sleep arose because beneficial effects of sleep were present after restudy cycles but were absent after retrieval cycles. This consistency in results across experiments indicates that the results of Experiment 1 were not due to ceiling effects on mean recall levels in the retrieval practice condition and that sleep reduces the testing effect regardless of the interference level at test. Similar to Experiment 1, the findings are more consistent with the bifurcation model than with the elaborative retrieval hypothesis.

Experiment 3

Although the results of Experiments 1 and 2 consistently indicate that sleep may reduce or even eliminate the testing effect, it could be argued that the usage of categorized item lists as study material in the two experiments deviates from typical testing effect studies that often used lists of unrelated paired associates as study material. Moreover, because previous research suggests that difficulty during retrieval practice can promote the testing effect (e.g., Kornell et al., 2011; Pyc & Rawson, 2009) and semantically

categorized material should be easier to practice than unrelated paired associates, it is unclear whether the results of Experiments 1 and 2 were affected by choice of study material and generalize to other material. Our goal in Experiment 3 was to address the issue.

Experiment 3 was similar to Experiment 1 but used a list of unrelated paired associates instead of a list of categorized items as study material. Like in Experiment 1, there were restudy and retrieval practice conditions. Subjects in the restudy condition were asked to restudy half of the initially studied paired associates once and the other half twice; similarly, in the retrieval-practice condition, subjects were asked to practice retrieval of half of the paired associates once and the other half twice. The final test was conducted after a 12-hr delay that included either regular sleep or wakefulness. Subjects were presented the stimulus words of the paired associates and were asked to recall the appropriate response words. The results of the experiment will help to clarify whether the reduction of the testing effect after sleep, as found in Experiments 1 and 2, is restricted to categorized study material or generalizes to paired associates.

Method

Participants. The final sample comprised 216 healthy students from Regensburg University ($M = 22.0$ years; range: 18–30 years; 31 male; seven additionally tested participants were excluded prior to data analysis because they had reported either alcohol intake or daytime napping between sessions). All participants were native German speakers and were distributed equally across conditions ($n = 36$ in each of the six conditions).

Material. Thirty-two unrelated neutral one- and two-syllable words were drawn from different semantic categories (Van Overschelde et al., 2004). Sixteen of these items were randomly chosen as stimulus words. The remaining 16 items were used for a second list of response items. A list of paired associates was created by randomly pairing the stimulus list with the response list. Eight of the paired associates were repeated once during retrieval practice or restudy, whereas the remaining paired associates were repeated twice; sets of paired associates were distributed across practice levels in a balanced manner.

Design. The experiment had the same 2 × 2 × 2 mixed-factorial design as Experiment 1. The factors of type of practice (restudy, retrieval practice) and delay (12-hr wake, 12-hr sleep) were again manipulated between subjects; the factor of practice level (low, high) was again manipulated within subjects. Like in Experiment 1, a short-delay condition was included to control for potential circadian effects. Half of the subjects in this condition participated at 9 a.m., the other half at 9 p.m.

Procedure.

Study and practice phase. Experiment 3 was conducted in a similar fashion as Experiment 1. During study, paired associates were presented successively in a random order, at a presentation rate of 5 s each. After one initial study cycle, one or two additional cycles consisted of either restudy or retrieval-practice trials, depending on practice condition. In restudy conditions, subjects were again presented with the paired associates for 5 s each. Half of the paired associates were restudied once (SS), whereas the other half of paired associates were restudied twice (SSS). In retrieval conditions, subjects were provided with the stimulus words and word stems of the response items for up to 7 s each and were asked to recall the corresponding items (M response time = 2.3 s, SD = 0.51). Retrieval practice was conducted in a covert fashion to allow for the simultaneous testing of two subjects. Instead of answering verbally during retrieval practice, participants were asked to press a corresponding key, thereby indicating whether they successfully retrieved the response item and triggering presentation of the next retrieval cue; covert retrieval has been shown to lead to testing effects similar to those of overt retrieval (Putnam & Roediger, 2013; Smith, Roediger, & Karpicke, 2013). Retrieval of the response item of half of the paired associates was practiced once (ST), whereas retrieval of the other half was practiced twice (STT). In both restudy and retrieval conditions, order of the paired associates was random.

Participants spent the interval between learning phase and test phase in parallel to Experiment 1. Consistent with Experiments 1 and 2, subjects in the 12-hr sleep conditions reported to have slept regularly during the night ($M = 7.4$ hr, $SD = 1.07$), whereas subjects in the 12-hr wake conditions reported not to have taken naps during the day. None of the subjects reported alcohol intake between sessions.

Test phase. At test, participants were presented with the stimulus words and initial letters of the response items of all 16 studied paired associates for 7 s each and were asked to recall the appropriate response item. Order of paired associates was random.

Results

Success rates during retrieval-practice cycles. A 2 × 2 × 2 ANOVA with the factors of time of day (9 a.m., 9 p.m.), retrieval practice (ST, STT), and delay (short delay, 12-hr delay) revealed that retrieval success was higher after two than after one retrieval-practice cycle (86.5% vs. 80.8%), $F(1, 104) = 11.13$, $MSE = 137.28$, $p = .001$, $\eta^2 = .10$, as indicated by subjects' keypresses. There were no other main effects and no interactions, ps $> .10$.

Final test (12-hr delay conditions). Figure 5 shows recall performance after the 12-hr delay. A 2 × 2 × 2 ANOVA with the factors of type of practice (restudy, retrieval practice), delay (12-hr wake, 12-hr sleep), and practice level (low, high) revealed significant main effects of delay, $F(1, 140) = 9.26$, $MSE = 673.69$, $p = .003$, $\eta^2 = .06$, and practice level, $F(1, 140) = 38.89$, $MSE = 235.77$, $p < .001$, $\eta^2 = .22$, but no main effect of type of practice, $F(1, 140) = 0.89$, $MSE = 673.69$, $p = .348$, $\eta^2 = .01$. The main effect of delay reflects that recall rates were overall higher in the 12-hr sleep condition than in the 12-hr wake condition (67.3% vs. 58.0%), whereas the main effect of practice level indicates better recall after the high than the low practice level (68.3% vs. 57.0%). More important, although all other interactions were nonsignificant (all ps $> .10$), there was a reliable two-way interaction between type of practice and delay, $F(1, 140) = 9.96$, $MSE = 673.69$, $p = .002$, $\eta^2 = .07$, indicating that the difference in memory performance after retrieval practice in comparison to restudy was modulated by delay condition. Consistently, planned comparisons showed significant testing effects after the 12-hr wake delay, for both the lower practice level (45.3% vs. 58.0%), $t(70) = 2.25$, $p = .028$, $d = 0.53$, and the higher practice level (58.1% vs. 70.5%), $t(70) = 2.94$, $p = .005$, $d = 0.69$, whereas no testing effects arose after sleep—and were even numerically reversed—for both the lower practice level (64.6% vs. 60.1%),

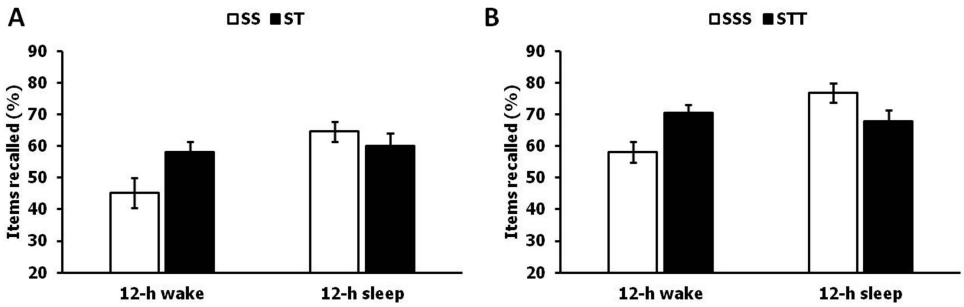

Figure 5. Mean recall performance on the final test of Experiment 3 as a function of delay (12-hr wake, 12-hr sleep) and type of practice (restudy, retrieval practice), separately for the low (A) and high (B) practice levels. Condition labels indicate study (S) and retrieval practice (T) cycles. Error bars represent standard errors.

$t(70) = 0.85$, $p = .399$, $d = 0.20$, and the higher practice level (76.7% vs. 67.7%), $t(70) = 1.89$, $p = .064$, $d = 0.44$. Further planned comparisons revealed that there was a sleep benefit for items that had been restudied once (SS), $t(70) = 3.33$, $p = .001$, $d = 0.79$, or twice (SSS), $t(70) = 4.09$, $p < .001$, $d = 0.96$, whereas there was no such benefit for items that had been subject to retrieval practice once (ST), $t(70) = 0.40$, $p = .688$, $d = 0.09$, or twice (STT), $t(70) = 0.62$, $p = .536$, $d = 0.15$.

Circadian control (short-delay condition). Table 2 shows mean recall levels after the short delay. A $2 \times 2 \times 2$ ANOVA with the factors of type of practice (restudy, retrieval practice), time of day (9 a.m., 9 p.m.), and practice level (low, high) revealed a significant main effect of practice level, $F(1, 68) = 20.49$, $MSE = 111.88$, $p < .001$, $\eta^2 = .23$, with items of the higher practice level being recalled better than items of the lower practice level (85.2% vs. 77.3%). No other effects emerged, indicating that there was no testing effect after short retention interval and that recall was unaffected by circadian effects, all ps $> .10$.

Additional analyses. Experiments 1 and 3 differ mainly in the study material used, with Experiment 1 using categorized item material and Experiment 3 using paired associates. Usage of paired associates induced a more difficult retrieval task than did usage of categorized items, as is reflected in the difference in mean success rates during retrieval practice between experiments (Experiment 1: 92.6%; Experiment 3: 83.7%), $t(214) = 7.32$, $p < .001$, $d = 1.00$. Numerically, this difference was also reflected in the size of the test–delay interaction. Indeed, averaged across the two practice levels, categorized lists showed a recall reduction from short delay to 12-hr wake delay of 16.9% after restudy and 7.1% after retrieval, whereas paired associates showed a recall reduction from short delay to 12-hr wake delay of 33.7% after restudy and 12.7% after retrieval. The results of a $2 \times 2 \times 2$ ANOVA with the factors of type of practice (restudy, retrieval practice), delay (9 a.m. short delay, 12-hr wake), and experiment (Experiment 1, Experiment 3) confirmed the numerical indication of a test–delay interaction, $F(1, 208) = 11.56$, $MSE = 241.10$, $p = .001$, $\eta^2 = .05$, showing more forgetting after restudy than after retrieval practice cycles; the numerically larger interaction in Experiment 3 than Experiment 1 did not reach significance, $F(1, 208) = 1.57$, $MSE = 241.10$, $p = .212$, $\eta^2 = .01$. Finally, experiment affected delay-induced forgetting, with a larger amount of forgetting in Experiment 3 than in Experiment 1, $F(1, 208) = 6.36$, $MSE = 241.10$, $p = .012$, $\eta^2 = .03$.

Table 2
Mean Recall Performance in the Short-Delay Condition of Experiment 3 as a Function of Time of Day, Type of Practice, and Practice Level

	Restudy				Retrieval practice			
	SS		SSS		ST		STT	
Time of day	*M*	*SE*	*M*	*SE*	*M*	*SE*	*M*	*SE*
9 a.m.	81.9	5.5	88.9	3.8	71.5	6.8	82.6	3.5
9 p.m.	82.6	5.2	86.8	4.9	72.9	5.6	82.6	4.9
Combined	82.3	3.7	87.8	3.0	72.2	4.3	82.6	3.0

Note. Condition labels indicate study (S) and retrieval-practice (T) cycles.

Discussion

As expected, usage of paired associates as study material rather than categorized items induced a more difficult retrieval task. Despite the resulting difference in mean success rates during retrieval practice, however, largely the same results arose as in Experiment 1. The results showed reliable testing effects after wake delay, after both one and two practice cycles, but no longer showed testing effects after sleep. In particular, again, sleep proved to be beneficial for restudied items but did not improve recall of retrieved items. Supporting the picture of the many parallels between experiments, test–delay interactions arose in both experiments and did not vary reliably in size. The finding of a reduced sleep effect on recall of retrieved items compared with restudied items is again consistent with the bifurcation model.

Experiments 4A and 4B

Because in Experiments 1–3, retrieval practice always induced higher mean recall levels after the wake delay than restudy did, in Experiments 4A and 4B, a greater number of restudy than retrieval practice cycles was used to rule out that the finding of sleep benefits for restudied but not for retrieved items was affected by differences in mean recall level. In Experiment 4A, a categorized item list was used as study material, whereas educational text material was used in Experiment 4B. In both experiments, following initial study, subjects received either three restudy cycles or

one retrieval-practice cycle. After a 12-hr delay interval that included either regular sleep or wakefulness, subjects were then asked to recall as many of the originally studied items (Experiment 4A) or as many details of the originally studied text (Experiment 4B) as possible. No short-delay conditions were included. On the basis of the results of Experiments 1–3, we expected in both experiments to find similar mean recall levels after wake delay in the two practice conditions. More important, we expected that sleep would again improve recall of the restudied material but leave recall of the retrieved material largely unaffected.

Experiment 4A

Method

Participants. The final sample consisted of 72 healthy students from Regensburg University (M = 22.9 years; range: 19–29 years; 16 male; three additionally tested subjects were eliminated from the sample prior to data analysis because they had reported alcohol consumption or daytime napping between sessions). All participants were native German speakers and were distributed equally across conditions (n = 18 in each of the four conditions).[3]

Material. A new categorized item list was constructed as study material. Thirty-six new concrete German nouns from six different semantic categories were selected using the same selection criteria as applied in Experiments 1 and 2 (Scheithe & Bäuml, 1995; Van Overschelde et al., 2004).

Design. The experiment had a 2 × 2 between-subjects design with the factors of practice type (restudy, retrieval practice) and delay (12-hr wake, 12-hr sleep). After initial encoding, half of the subjects restudied the items of all six categories three times (SSSS), whereas the other half of subjects practiced all items' retrieval once (ST). The 12-hr wake and the 12-hr sleep conditions were conducted in an identical manner as in Experiments 1 and 2. Subjects started the experiment at 9 a.m. or 9 p.m., respectively.

Procedure. The initial study cycle for all items was carried out in parallel to Experiment 1 and was followed by either three additional study cycles or one retrieval-practice cycle. Like in Experiment 3, retrieval practice was conducted covertly (M response time = 1.9 s, SD = 0.48). The final test was again conducted in parallel to Experiment 1.

Regarding compliance with instructions, subjects in the 12-hr sleep conditions reported to have slept regularly during the night (M = 7.5 hr, SD = 0.88), whereas subjects in the 12-hr wake conditions reported that they did not take naps during the day. None of the subjects reported alcohol intake between sessions.

Results

Success rates during retrieval-practice trials. A comparison between the 12-hr wake and the 12-hr sleep conditions revealed no difference regarding success rates during retrieval practice (86.7% vs. 86.1%), $t(34) = 0.26$, $p = .793$, $d = 0.09$, as they were indicated by subjects' keypresses.

Final test (12-hr delay conditions). Figure 6A shows recall performance after the 12-hr delay. A 2 × 2 ANOVA with the factors of practice type (restudy, retrieval practice) and delay (12-hr wake, 12-hr sleep) revealed a significant main effect of practice type, $F(1, 68) = 8.70$, $MSE = 104.32$, $p = .004$, $\eta^2 = .11$, with better recall in the restudy than in the retrieval-practice condition (75.5% vs. 68.4%). Additionally, there was a significant main effect of delay, $F(1, 68) = 11.98$, $MSE = 104.32$, $p < .001$, $\eta^2 = .15$, reflecting better overall recall in the 12-hr sleep condition than in the 12-hr wake condition (76.2% vs. 67.8%). Most important, there was a significant interaction between the two factors, $F(1, 68) = 7.60$, $MSE = 104.32$, $p = .008$, $\eta^2 = .10$. Planned comparisons showed that there was no significant difference in recall performance between the restudy condition (SSSS) and the retrieval practice condition after the wake delay (ST; 68.1% vs. 67.6%), $t(34) = 0.12$, $p = .903$, $d = 0.04$. However, after the sleep delay, recall was significantly higher after restudy (SSSS) than after retrieval practice (ST; 83.0% vs. 69.3%), $t(34) = 4.58$, $p < .001$, $d = 1.53$. Consistently, there was a significant sleep effect in the restudy condition (SSSS), $t(34) = 4.29$, $p < .001$, $d = 1.43$, but not in the retrieval-practice condition (ST), $t(34) = 0.512$, $p = .612$, $d = 0.17$.

Discussion

As expected, after the wake delay, mean recall levels were comparable between the retrieval-practice (ST) and the restudy (SSSS) conditions. Thus, if mean recall level were critical for the sleep effect, then similar effects of sleep should have arisen in the SSSS and ST conditions. The results turned out otherwise, however. Despite the equivalence in wake mean recall level, sleep affected retrieved and restudied items differently, improving memories after three restudy cycles but not after the single retrieval practice cycle. This finding mimics the results of Experiments 1–3, which also found sleep benefits to be restricted to restudy conditions and to not generalize to retrieval-practice conditions. Experiment 4B was conducted to replicate the results of Experiment 4A and examine whether they generalize to educational text material. To address the issue, in Experiment 4B, we used text material as it was used in the study by Roediger and Karpicke (2006).

Experiment 4B

Method

Participants. Ninety-six healthy subjects participated in the experiment (M = 23.4 years; range: 18–30 years; 39 male; five additional subjects were eliminated from the sample prior to data analysis because of alcohol consumption or daytime napping between sessions). Participants were native German speakers and were distributed equally across conditions (n = 24 in each of the four conditions).

Material. Subjects studied one of two text passages, each covering a specific topic ("The Sun" or "Sea Otters"; see Roediger & Karpicke, 2006). Scoring of recall performance was conducted using German translations of 30 idea units for each passage. Word length of the two passages was comparable (242 vs. 254 words).

Design. A 2 × 2 between-subjects design with the factors of type of practice (restudy, retrieval practice) and delay (12-hr wake,

[3] In this experiment, there was only one practice level in the restudy and retrieval conditions, meaning that more data per subject were collected for the two types of items than in the previous experiments. We therefore reduced number of subjects for each of the four experimental conditions.

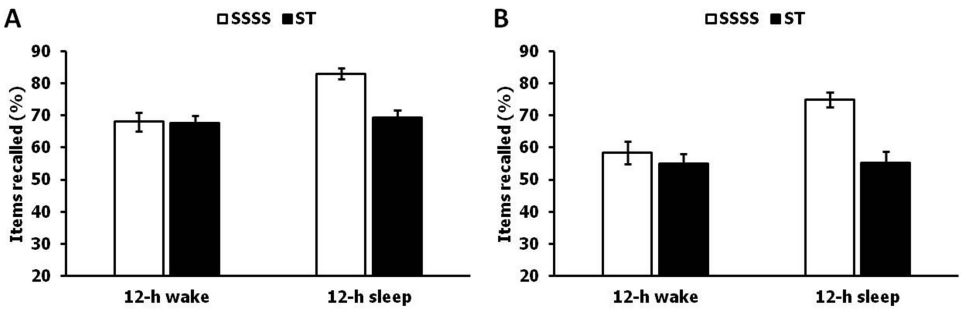

Figure 6. Mean recall performance on the final test of Experiment 4A (A) and Experiment 4B (B) as a function of delay (12-hr wake, 12-hr sleep) and type of practice (restudy, retrieval practice). Condition labels indicate study (S) and retrieval practice (T) cycles. Error bars represent standard errors.

12-hr sleep) was used. After initial study, half of the subjects restudied the text passage three times (SSSS; restudy condition), whereas the other half of the subjects practiced its retrieval once (ST; retrieval-practice condition). The 12-hr wake and the 12-hr sleep conditions were identical to those of the previous experiments and started at 9 a.m. or 9 p.m. The text passages were used equally often in each condition.

Procedure. Study and test procedures closely followed those of Experiment 2 of Roediger and Karpicke (2006). Participants were asked to study the respective text passage in an initial study cycle (5 min) that was followed by three additional study cycles of the same kind (SSSS) or one retrieval practice cycle (10 min; ST). During study cycles, subjects were asked to memorize all of the details from the passage, whereas during retrieval practice cycles, they were provided with a sheet and the title of the passage and asked to write down everything they could remember, without concern for correct wording or order. Between restudy trials, participants solved arithmetic problems (2 min). Study phase was followed by a short distractor phase before subjects were dismissed from the first session. During the second session, all subjects were engaged in the final test that was conducted in exactly the same way as during the retrieval-practice phase.

Like in the previous experiments, subjects in the 12-hr sleep conditions reported to have slept regularly during the night (M = 7.7 hr, SD = 1.06), whereas subjects in the 12-hr wake conditions reported not to have taken naps during the day. None of the subjects reported alcohol intake between sessions.

Results

Success rates during retrieval-practice cycles. The 12-hr wake and 12-hr sleep conditions did not differ in success rates during retrieval practice (58.8% vs. 57.9%), $t(46) = 0.21$, $p = .834$, $d = 0.06$.

Final test. Figure 6B shows recall performance after the 12-hr delay. A 2 × 2 ANOVA with the factors of type of practice (restudy, retrieval practice) and delay (12-hr wake, 12-hr sleep) was conducted. It revealed a main effect of type of practice, $F(1, 92) = 10.69$, $MSE = 225.08$, $p = .002$, $\eta^2 = .10$, with better recall in the restudy than in the retrieval-practice condition (66.7% vs. 56.7%). Additionally, there was a main effect of delay, $F(1, 92) = 10.10$, $MSE = 225.08$, $p < .002$, $\eta^2 = .10$, indicating better overall recall in the 12-hr sleep condition than in the 12-hr wake condition (66.5% vs. 56.8%). Most important, there was a significant interaction between the two factors, $F(1, 92) = 4.70$, $MSE = 225.08$, $p = .033$, $\eta^2 = .05$. In fact, planned comparisons showed that there was no significant difference in recall performance between restudy (SSSS) and retrieval-practice (ST) conditions after the wake delay (58.5% vs. 55.1%), $t(46) = 0.75$, $p = .458$, $d = 0.22$, whereas after the sleep delay, recall was significantly higher after restudy (SSSS) than retrieval practice (ST; 74.9% vs. 58.2%), $t(46) = 4.02$, $p < .001$, $d = 1.16$. Consistently, there was again a significant sleep effect in the restudy condition (SSSS), $t(46) = 3.91$, $p < .001$, $d = 1.13$, but not in the retrieval-practice condition (ST), $t(46) = 0.692$, $p = .492$, $d = 0.20$.

Discussion

Experiment 4B induced a markedly more difficult retrieval task than did Experiments 1–4A, as is reflected in the difference in mean success rates during retrieval practice between experiments. Despite the more demanding retrieval task, however, the effects of sleep on restudied and retrieved items mimicked those found in the previous experiments. Again, sleep improved recall of the restudied items but left recall of the retrieved items unaffected. This finding arose even though restudy was conducted over two additional practice cycles and thus led to similar mean recall levels after the 12-hr wake delay and the retrieval practice conditions, which indicates that the previous findings were not driven by differences in mean recall level between restudied and retrieved items. Empirically, the results generalize the main findings of Experiments 1–4A to educational text material. Theoretically, they again support the bifurcation model.

General Discussion

Our goal in the present series of experiments was to investigate possible differential effects of wake and sleep delay on the testing effect. Using categorized word lists (Experiments 1, 2, and 4A), lists of paired associates (Experiment 3), and educational text material (Experiment 4B), we found that the results of the experiments consistently demonstrated that sleep can influence the testing effect. Both after one and after two practice cycles, Experiments 1–3 showed typical testing effects after a 12-hr wake delay, that is, better recall after retrieval than restudy trials, which replicates the prior work (e.g., Hogan & Kintsch, 1971; Roediger &

Karpicke, 2006). Going beyond the prior work, however, the same experiments showed reduced testing effects (Experiment 2) or even eliminated testing effects (Experiments 1 and 3) after a 12-hr delay that included nocturnal sleep.

The results of Experiments 1–4 indicate that the reduction of the testing effect after sleep arises because sleep is beneficial for recall after restudy trials but is not beneficial for recall after retrieval-practice trials. Indeed, sleep benefited recall in all eight restudy conditions of the present experiments, whereas in all eight retrieval-practice conditions, no reliable benefit of sleep arose. The absence of beneficial effects of sleep on recall of retrieved items was observed when recall levels were relatively high (Experiments 1 and 4A) and when recall levels were reduced and far from ceiling (Experiments 2, 3, and 4B); the same pattern arose when the retrieval task was relatively difficult (Experiment 4B), moderately difficult (Experiment 3), and relatively easy (Experiments 1, 2, and 4A), and it was present both when wake mean recall levels were higher for retrieved than restudied items (Experiments 1–3) and when wake mean recall levels were equated (Experiments 4A and 4B). Together, the results suggest that prior retrieval practice can make recall relatively immune to the beneficial effects of sleep, a finding that is at the core of the observed reduction of the testing effect after sleep.

The results of the present experiments are well in line with the bifurcation model (Halamish & Bjork, 2011; Kornell et al., 2011). This model assumes that items that are successfully retrieved during retrieval practice (but not the nonretrieved items) are not only strengthened but strengthened to a higher degree than restudied items are. This high degree of strengthening may enable many of the successfully retrieved items to remain above recall threshold not only after short delay but also after prolonged wake delay, thus leaving not much room for additional beneficial effects of sleep on recall performance. In contrast, because restudy supposedly strengthens repeated items less effectively than retrieval practice does, sleep may help some of the restudied items to cross recall threshold, thus improving recall chances of the restudied items after sleep. If so, the testing effect may be present after waking but be reduced or even absent after sleep, which is exactly what the present results show. The bifurcation model also predicts that the testing effect should increase with length of (the waking) retention interval (Kornell et al., 2011) and be larger in the presence than in the absence of retroactive interference (Halamish & Bjork, 2011). The short-delay and 12-hr wake delay conditions of Experiments 1 and 3 confirm the first prediction and the results of Experiments 1 and 2 confirm the second. The present results thus show a high degree of consistency with the bifurcation model.

The present results appear less well consistent with the elaborative retrieval hypothesis. According to this hypothesis, the testing effect arises because retrieval practice leads to more elaborative processing than restudy does, fostering the activation of extra semantic information that may help successfully recalling retrieved items on a later memory test (e.g., Carpenter, 2009; Pyc & Rawson, 2010). On the basis of this account, one may expect sleep to maintain or even increase the testing effect, because sleep has repeatedly been found to enhance memory for semantic information (e.g., Cai et al., 2009; McKeon et al., 2012; Payne et al., 2009). The present finding of a reduction or even elimination of the testing effect after sleep and no sleep benefit for retrieved items is not easily reconciled with this expectation, indicating that extra processing of semantic information during retrieval cycles may not have induced the testing effect in the present experiments.[4]

The experiments in this study differ in retrieval difficulty during retrieval practice. Whereas Experiments 1, 2, and 4A used categorized item lists and induced relatively easy retrieval tasks, Experiment 3 used lists of paired associates and Experiment 4B used educational text material, thus inducing more difficult retrieval tasks. Variations in the difficulty of the retrieval-practice task can be of relevance for the bifurcation model. The reason is that variation in retrieval difficulty induces variation in the degree to which the strength distribution of retrieved and nonretrieved items in the retrieval practice condition is bifurcated. Indeed, more difficult retrieval tasks, like the one used in Experiment 4B, should induce a higher degree of bifurcation than easier retrieval tasks do, like the one used in Experiment 1. Although a higher degree of bifurcation can increase the size of the testing effect (Kornell et al., 2011), there is no simple relationship between retrieval difficulty and size of the testing effect in the bifurcation model. In fact, even low levels of bifurcation can result in robust testing effects if almost all retrieved items are above threshold and the retrieved items are farther above threshold than are the restudied items—conditions likely met in Experiments 1 and 2. In such case, when the retention interval is increased, restudied items begin to cross below threshold before the retrieved items do so, leading to the typical testing effect finding (see Kornell et al., 2011, p. 89). The present finding that the testing effect and the effect of sleep on retrieved items did not vary much with retrieval difficulty in the retrieval practice phase thus is in line with the bifurcation model.

Prior work on the effects of sleep on memory performance showed that sleep does not benefit all memories equally (for a review, see Stickgold & Walker, 2013). For instance, emotional memories have been found to show more sleep benefits than neutral memories do (Payne et al., 2008), and memories considered relevant for the future show more sleep benefits than supposedly irrelevant material does (Wilhelm et al., 2011). In all of these cases, the proposal has been that sleep does not treat all memories equally and strengthens some memories more than others. The present results are basically consistent with such a proposal by demonstrating that sleep benefits recall of restudied items much more than it benefits recall of retrieved items. However, the present results are not necessarily inconsistent with the alternative view that sleep treats retrieved and restudied items equally. Indeed, following the bifurcation model, retrieved items may already be strengthened to such a high degree that an additional strengthening effect of sleep on the retrieved items may not easily be detected in typical testing effect experiments.

The present results on the effects of sleep for restudied items may provide some support for the latter view. In the present experiments, beneficial effects of sleep for restudied items were observed for all types of materials that were used in this study, and they were largely unaffected by the number of restudy cycles. In fact, averaged across experiments, the size of the observed sleep benefit for restudied items was about the same regardless of

[4] Although elaborative retrieval and the bifurcation model differ in the degree to which they can explain the present results and generally may lead to different predictions on the role of sleep for the testing effect, it should again be emphasized that the two accounts are not mutually exclusive and might be at work simultaneously (see also Kornell et al., 2011).

whether there were one restudy cycle (sleep benefit of 14.7%), two restudy cycles (sleep benefit of 15.3%), or three restudy cycles (sleep benefit of 16.4%), suggesting that sleep strengthens memories largely independent of their original strength level. If retrieval practice differed from restudy only by increasing the strength level of practiced memories even farther, as is suggested by the bifurcation model, then arguably sleep may strengthen retrieved items in a manner very similar to how it strengthens restudied items. Further work is required to address the issue in more detail. Such work may use retention intervals of several days or even weeks, so a larger proportion of the retrieved items can cross below recall threshold and possible sleep benefits on recall of retrieved items become more easily visible. The results of such work would lead to further tests of the bifurcation model as well as the selective sleep benefits proposal, providing more detailed information about the degree to which sleep strengthens restudied and retrieved items.

A review of the testing effect literature shows that typical testing effect studies differ in whether feedback is provided after retrieval practice (see Roediger & Butler, 2011). The primary goal of the bifurcation model of the testing effect has been to capture direct effects of testing, that is, effects that arise through the act of retrieval itself in the absence of feedback. The present experiments meet these conditions and the results fit well with the bifurcation model. However, other research documents additional mediated effects of testing, for instance, showing that retrieval practice can facilitate subsequent restudy when feedback is provided (see Arnold & McDermott, 2013a, 2013b). In particular, feedback may play a critical role for retrieval-induced elaboration processes, for instance, enabling the generation of more potent semantic mediators between cue and target during paired associate learning (Pyc & Rawson, 2010, 2012). If elaborative processes were indeed more important for mediated than direct effects of testing, then sleep might influence mediated effects of testing differently than direct effects of testing, possibly maintaining or even enhancing the mediated effects. Future work may therefore investigate the role of feedback for the influence of sleep on the testing effect. Such work may provide information about the role of sleep for direct versus mediated effects of testing, as well as the contribution of retrieval-induced strengthening versus elaborative processes for the testing effect.

Finally, although the present findings are consistent with the bifurcation model and the view that sleep can reduce the testing effect, they are also in line with an intriguing alternative view of the testing effect.[5] This view is largely based on three findings: (a) The testing effect is often absent after a short wake delay but is present after a longer wake delay (see Experiments 1 and 3), (b) the testing effect is larger in the presence than the absence of interference at test (see Experiments 1 and 2), and (c) the pattern of results after a short wake delay is very similar to the patterns observed after a sleep delay (see Experiments 1 and 3). Indeed, on the basis of these findings, the view may arise that sleep is not necessary to reduce the testing effect but rather prolonged wakefulness is the key to increasing the testing effect. According to such view, longer wake intervals may increase (extraexperimental) interference at test and thus increase the testing effect (Halamish & Bjork, 2011; see Experiment 2), whereas no such increase in interference may arise during sleep intervals, thus leading to similar results after short wake and sleep delays. The present study was not designed to evaluate this alternative view of the testing effect; thus, future work is required to address the issue in more detail.

Conclusions

Using a wide range of study materials, the present research demonstrates in a series of five experiments that sleep can reduce and sometimes even eliminates the testing effect. At the core of this finding is the result that sleep benefits recall of restudied items but leaves recall of retrieved items unaffected. The results arose regardless of mean recall level, difficulty of retrieval task, and interference level at test. The finding is consistent with the view that retrieval strengthens items to a higher degree than restudy does so that many of the successfully retrieved items remain above recall threshold, even after prolonged wake delay, and additional sleep-induced strengthening may not improve recall of retrieved items any further.

[5] This view was suggested to us by Michael Scullin during the review process.

References

Abel, M., & Bäuml, K.-H. T. (2012). Retrieval-induced forgetting, delay, and sleep. *Memory, 20,* 420–428. doi:10.1080/09658211.2012.671832

Abel, M., & Bäuml, K.-H. T. (2013). Sleep can eliminate list-method directed forgetting. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 39,* 946–952. doi:10.1037/a0030529

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

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

Arnold, K. M., & McDermott, K. B. (2013a). Free recall enhances subsequent learning. *Psychonomic Bulletin & Review, 20,* 507–513. doi:10.3758/s13423-012-0370-3

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

Barrett, T. R., & Ekstrand, B. R. (1972). Effect of sleep on memory: III. Controlling for time-of-day effects. *Journal of Experimental Psychology, 96,* 321–327. doi:10.1037/h0033625

Bäuml, K.-H. T., Pastötter, B., & Hanslmayr, S. (2010). Binding and inhibition in episodic memory: Cognitive, emotional, and neural processes. *Neuroscience and Biobehavioral Reviews, 34,* 1047–1054. doi:10.1016/j.neubiorev.2009.04.005

Cai, D. J., Mednick, S. A., Harrison, E. M., Kanady, J. C., & Mednick, S. C. (2009). REM, not incubation, improves creativity by priming associative networks. *PNAS: Proceedings of the National Academy of Sciences, USA, 106,* 10130–10134. doi:10.1073/pnas.0900271106

Carpenter, S. K. (2009). Cue strength as a moderator of the testing effect: The benefits of elaborative retrieval. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 35,* 1563–1569. doi:10.1037/a0017021

Carpenter, S. K. (2011). Semantic information activated during retrieval contributes to later retention: Support for the mediator effectiveness hypothesis of the testing effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 37,* 1547–1552. doi:10.1037/a0024140

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

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

Darsaud, A., Dehon, H., Lahl, O., Sterpenich, V., Boly, M., Dang-Vu, T., Desseilles, M., . . . Collette, F. (2011). Does sleep promote false memories? *Journal of Cognitive Neuroscience, 23,* 26–40. doi:10.1162/jocn.2010.21448

Diekelmann, S., & Born, J. (2010). The memory function of sleep. *Nature Reviews Neuroscience, 11,* 114–126.

Fenn, K. M., Gallo, D. A., Margoliash, D., Roediger, H. L., III, & Nusbaum, H. C. (2009). Reduced false memory after sleep. *Learning & Memory, 16,* 509–513. doi:10.1101/lm.1500808

Ficca, G., Lombardo, P., Rossi, L., & Salzarulo, P. (2000). Morning recall of verbal material depends on prior sleep organization. *Behavioural Brain Research, 112,* 159–163. doi:10.1016/S0166-4328(00)00177-7

Gais, S., Lucas, B., & Born, J. (2006). Sleep after learning aids memory recall. *Learning & Memory, 13,* 259–262. doi:10.1101/lm.132106

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. doi:10.1037/a0023219

Hogan, R. M., & Kintsch, W. (1971). Differential effects of study and test trials on long-term recognition and recall. *Journal of Verbal Learning and Verbal Behavior, 10,* 562–567. doi:10.1016/S0022-5371(71)80029-4

Jenkins, J. G., & Dallenbach, K. M. (1924). Obliviscence during sleep and waking. *The American Journal of Psychology, 35,* 605–612. doi:10.2307/1414040

Johns, M. W. (1991). A new method for measuring daytime sleepiness: The Epworth Sleepiness Scale. *Sleep, 14,* 540–545.

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. doi:10.1016/j.jml.2011.04.002

McDaniel, M. A., & Masson, M. E. J. (1985). Altering memory representations through retrieval. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 11,* 371–385. doi:10.1037/0278-7393.11.2.371

McKeon, S., Pace-Schott, E. F., & Spencer, R. M. C. (2012). Interaction of sleep and emotional content on the production of false memories. *PLOS One, 7,* Article e49353. doi:10.1371/journal.pone.0049353

Oswald, W. D., & Roth, E. (1987). *Der Zahlen-Verbindungs-Test (ZVT)* [Connect-the-Numbers Test]. Göttingen, Germany: Hogrefe.

Payne, J. D., Schacter, D. L., Propper, R. E., Huang, L.-W., Wamsley, E. J., Tucker, M. A., . . . Stickgold, R. (2009). The role of sleep in false memory formation. *Neurobiology of Learning and Memory, 92,* 327–334. doi:10.1016/j.nlm.2009.03.007

Payne, J. D., Stickgold, R., Swanberg, K., & Kensinger, E. A. (2008). Sleep preferentially enhances memory for emotional components of scenes. *Psychological Science, 19,* 781–788. doi:10.1111/j.1467-9280.2008.02157.x

Plihal, W., & Born, J. (1997). Effects of early and late nocturnal sleep on declarative and procedural memory. *Journal of Cognitive Neuroscience, 9,* 534–547. doi:10.1162/jocn.1997.9.4.534

Potts, R., & Shanks, D. R. (2012). Can testing immunize against interference? *Journal of Experimental Psychology: Learning, Memory, and Cognition, 38,* 1780–1785. doi:10.1037/a0028218

Putnam, A. L., & 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. (2009). Testing the retrieval effort hypothesis: Does greater difficulty correctly recalling information lead to higher levels of memory? *Journal of Memory and Language, 60,* 437–447. doi:10.1016/j.jml.2009.01.004

Pyc, M. A., & Rawson, K. A. (2010, October 15). Why testing improves memory: Mediator effectiveness hypothesis. *Science, 330,* 335. doi:10.1126/science.1191465

Pyc, M. A., & Rawson, K. A. (2012). Why is test–restudy practice beneficial for memory? An evaluation of the mediator shift hypothesis. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 38,* 737–746. doi:10.1037/a0026166

Raaijmakers, J. G., & Jakab, E. (2013). Rethinking inhibition theory: On the problematic status of the inhibitory theory for forgetting. *Journal of Memory and Language, 68,* 98–122. doi:10.1016/j.jml.2012.10.002

Racsmány, M., Conway, M. A., & Demeter, G. (2010). Consolidation of episodic memories during sleep: Long-term effects of retrieval practice. *Psychological Science, 21,* 80–85. doi:10.1177/0956797609354074

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, & McDermott, K. B. (1995). Creating false memories: Remembering words not presented in lists. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 21,* 803–814. doi:10.1037/0278-7393.21.4.803

Scheithe, K., & Bäuml, K.-H. (1995). Deutschsprachige Normen für Vertreter von 48 Kategorien [German language norms for representatives of 48 categories]. *Sprache & Kognition, 14,* 39–43.

Scullin, M. K., & McDaniel, M. A. (2010). Remembering to execute a goal: Sleep on it! *Psychological Science, 21,* 1028–1035. doi:10.1177/0956797610373373

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

Stickgold, R., & Walker, M. P. (2013). Sleep-dependent memory triage: Evolving generalization through selective processing. *Nature Neuroscience, 16,* 139–145. doi:10.1038/nn.3303

Szpunar, K. K., McDermott, K. B., & Roediger, H. L., III. (2007). Expectation of a final cumulative test enhances long-term retention. *Memory & Cognition, 35,* 1007–1013. doi:10.3758/BF03193473

Talamini, L. M., Nieuwenhuis, I. L., Takashima, A., & Jensen, O. (2008). Sleep directly following learning benefits consolidation of spatial associative memory. *Learning & Memory, 15,* 233–237. doi:10.1101/lm.771608

Toppino, T. C., & Cohen, M. S. (2009). The testing effect and the retention interval. *Experimental Psychology, 56,* 252–257. doi:10.1027/1618-3169.56.4.252

Tucker, M. A., & Fishbein, W. (2008). Enhancement of declarative memory performance following a daytime nap is contingent on strength of initial task acquisition. *Sleep, 31,* 197–203.

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

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. doi:10.1111/j.1467-9280.1992.tb00036.x

Wilhelm, I., Diekelmann, S., Molzow, I., Ayoub, A., Mölle, M., & Born, J. (2011). Sleep selectively enhances memory expected to be of future relevance. *The Journal of Neuroscience, 31,* 1563–1569. doi:10.1523/JNEUROSCI.3575-10.2011

Received December 18, 2013
Revision received April 7, 2014
Accepted April 8, 2014 ■