DOI 10.3758/s13415-016-0407-6

Reversing the testing effect by feedback: Behavioral and electrophysiological evidence

Bernhard Pastötter[1] · Karl-Heinz T. Bäuml[1]

Published online: 8 February 2016

Abstract The testing effect refers to the finding that retrieval practice of previously studied information enhances its long-term retention more than restudy practice does. Recent work showed that the testing effect can be dramatically reversed when feedback is provided to participants during final recall testing (Storm, Friedman, Murayama, & Bjork, 2014). Following this prior work, in this study, we examined the reversal of the testing effect by investigating oscillatory brain activity during final recall testing. Twenty-six healthy participants learned cue–target word pairs and underwent a practice phase in which half of the items were retrieval practiced and half were restudy practiced. Two days later, two cued recall tests were administered, and immediate feedback was provided to participants in Test 1. Behavioral results replicated the prior work by showing a testing effect in Test 1, but a reversed testing effect in Test 2. Extending the prior work, EEG results revealed a feedback-related effect in alpha/lower-beta and retrieval-related effects in slow and fast theta power, with practice condition modulating the fast theta power effect for items that were not recalled in Test 1. The results indicate that the reversed testing effect can arise without differential strengthening of restudied and retrieval-practiced items via feedback learning. Theoretical implications of the findings, in particular with respect to the distribution-based bifurcation model of testing effects (Kornell, Bjork, & Garcia, 2011), are discussed.

Keywords EEG brain oscillations · Feedback learning · (Reversed) testing effect

Introduction

Testing and reversed testing effects

Retrieval practice can promote long-term memory. The most explored benefit of retrieval practice, which is referred to as the testing effect, is the finding that retrieval practice of previously studied information can improve its long-term retention more than restudy practice does (e.g., Hogan & Kintsch, 1971; Roediger & Karpicke, 2006, for a review, see Roediger & Butler, 2011). That is, when participants are asked to repeatedly retrieve or restudy previously studied items and long-term retention of items is assessed in a delayed criterion test, the retrieved items are typically better recalled than the restudied items. The testing effect is most prominent when retrieval practice is difficult but successful (Carpenter, 2009; Kornell, Bjork, & Garcia, 2011) and when the final criterion test is administered after a relatively long retention interval (Roediger & Karpicke, 2006; Toppino & Cohen, 2009). In addition to the testing effect, other benefits of retrieval practice can arise. For instance, retrieval practice in comparison to restudy practice can enhance transfer of learning (Butler, 2010; Carpenter & Kelly, 2012), potentiate new learning (Arnold & McDermott, 2013; Pastötter, Schicker, Niedernhuber, & Bäuml, 2011), and insulate tested material against interference from other information (Halamish & Bjork, 2011; Kliegl & Bäuml, 2016). On the basis of these and related findings, it has been argued that retrieval practice is a more effective learning strategy than restudy practice, in both the laboratory and the classroom (Dunlosky, Rawson, Marsh, Nathan, & Willingham, 2013; Roediger & Butler, 2011).

✉ Bernhard Pastötter
bernhard.pastoetter@psychologie.uni-regensburg.de

[1] Department of Experimental Psychology, Regensburg University, Universitätsstr. 31, 93053 Regensburg, Germany

Challenging this view, however, a recent study by Storm, Friedman, Murayama, and Bjork (2014, Experiment 1) provided evidence that the testing effect can be dramatically reversed when multiple recall tests are administered and feedback is provided to participants during final recall testing. In Session 1 of this study, participants studied 36 Swahili–English word pairs and repeatedly retrieved or restudied subsets of these items. In retrieval-practice trials, the Swahili cues were presented, and participants were instructed to retrieve the English targets (no feedback was provided during retrieval practice), whereas in restudy-practice trials, the Swahili–English word pairs were shown intact and participants were instructed to study the items once again. One week later, in Session 2, participants then were given six cued recall tests, in which the Swahili words were shown as cues and participants were instructed to retrieve the English targets. Importantly, immediate feedback was provided to participants in these cued recall tests, and participants after each recall trial were shown the intact word pair. This was done regardless of whether target recall was successful. The results were striking, showing a reliable testing effect (i.e., higher recall of retrieval-practiced than restudied items) in the first recall test (Test 1), but a reversed testing effect with higher recall of restudied than retrieval-practiced items in the second recall test (Test 2) and all subsequent recall tests. These results suggest that a single feedback-induced restudy opportunity during final recall testing can be sufficient to reverse the testing effect, such that restudied items become more recallable than retrieval-practiced items on subsequent recall tests.

Explanations of the testing effect and its reversal

Two recent process accounts of the testing effect are the semantic elaboration account and the episodic context account. The semantic elaboration account assumes that testing of previously studied information improves its long-term retention because retrieval practice, more than restudy practice, induces elaborative or deep processing of the information (Carpenter, 2009; McDaniel & Masson, 1985). The proposal is that, when participants attempt to retrieve a target item from memory during retrieval practice, semantically related nontarget information may be activated. When retrieval of the target item is successful, this nontarget information may become linked to the target information. In the final memory test, the nontarget information then can serve as a retrieval route to the target information and thus improve target retrieval. The semantic-elaboration account is supported by various findings from the testing-effect literature (e.g., Carpenter, 2009; Pyc & Rawson, 2009). In contrast to the semantic elaboration account, the episodic context account assumes that testing of previously studied information improves its long-term retention because retrieval practice can update context representations of the studied items (Karpicke, Lehman, & Aue, 2014). Specifically, the account assumes that during retrieval practice of items in a current context (context B), participants must reinstate the study context (context A) in order to recall the studied items. When retrieval of a target item is successful, features from the reinstated context A and the current context B may be combined to a composite context representation AB, which contains features that are associated with both contexts A and B. In the final memory test, reinstatement of features from either context A or context B then may serve as effective retrieval cues to evoke the target item from memory. Context updating via retrieval practice thus enhances memory search and improves target recall. The episodic context account can provide explanations for a number of findings in the retrieval-based learning literature (e.g., Kliegl & Bäuml 2016; Lehman, Smith, & Karpicke, 2014).

Importantly, while both the semantic elaboration account and the episodic context account provide explanations of the testing effect, they cannot easily explain the reversal of the testing effect as reported by Storm et al. (2014). However, another recently suggested explanation of testing effects, the distribution-based bifurcation model (Halamish & Bjork, 2011; Kornell et al., 2011) can account for both the testing effect and its reversal. The bifurcation model is a descriptive memory model or framework of testing effects, not a theory in the process-model sense. It incorporates four main assumptions (see Fig. 1). The first assumption is that initially studied items are normally distributed on some memory-strength dimension. The second assumption is that restudy practice strengthens items about equally, moving the item distribution of restudied items to the right. The third and core assumption is that retrieval practice creates a bifurcated item distribution, in which items that are successfully retrieved are strengthened to a higher degree than the restudied items, whereas items that are not successfully retrieved remain on their original memory-strength level. The fourth assumption finally is that, when the retention interval between retrieval or restudy practice and the final test is increased, all items are reduced in strength at a comparable rate, which moves the two item distributions back to the left. These assumptions are sufficient to explain both the testing effect in Test 1 and the reversed testing effect in Test 2 (and all subsequent recall tests) in Storm et al.'s (2014) study. In fact, by assuming that more of the restudied items than the successfully retrieved items fall below the recall threshold after delay, the model can explain the testing effect in Test 1 (see Fig. 1a). By assuming that more of the restudied items than the unsuccessfully retrieved items move over the recall threshold due to feedback learning in Test 1, the model can explain the reversed testing effect in Test 2 (see Fig. 1b). Notably, the model can also explain other recently reported effects, such as the influences of sleep and retroactive interference on the testing effect (e.g., Bäuml, Holterman, & Abel, 2014; Halamish & Bjork, 2011). Although it is a highly important research question which cognitive processes mediate

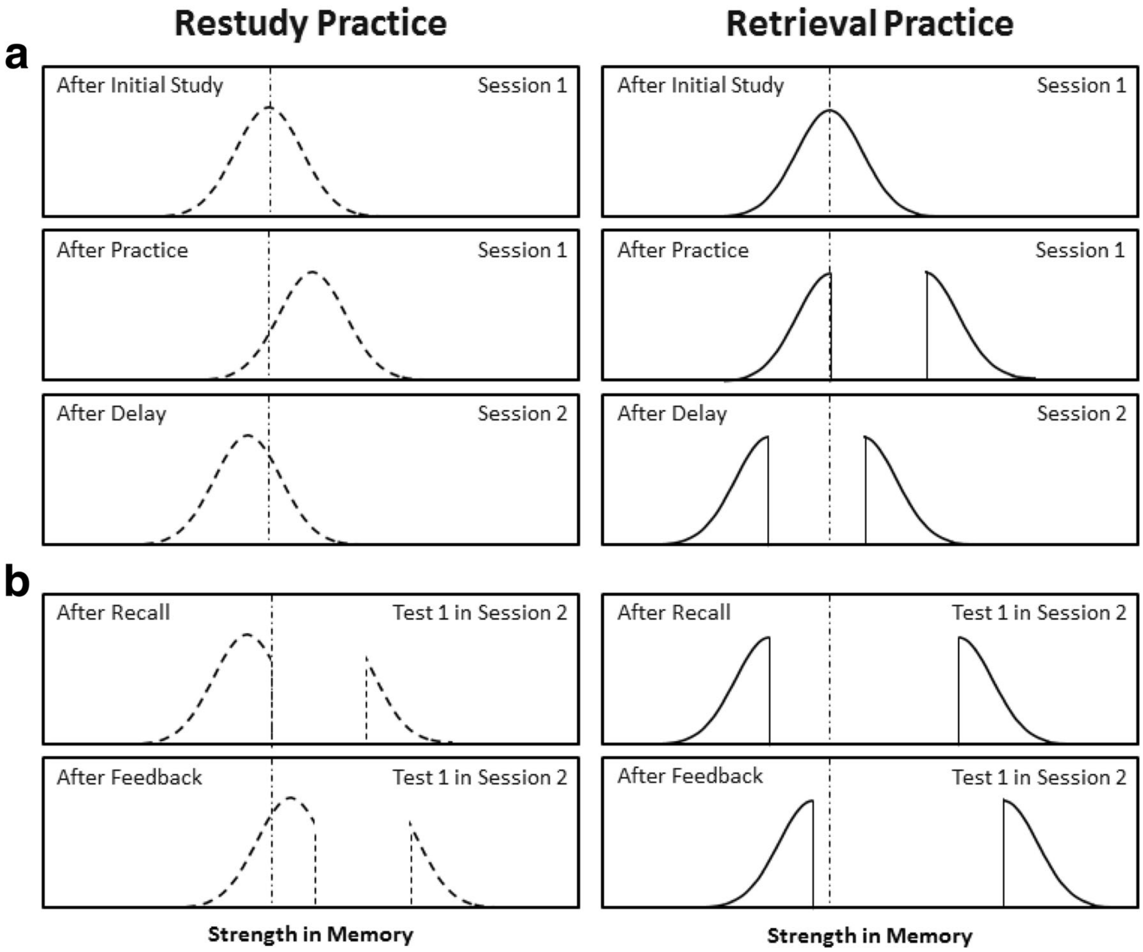

Fig. 1 **a** Illustration of the distribution-based bifurcation model with memory strength distributions of two hypothetical item sets (Kornell, Bjork, & Garcia, 2011). The left column shows items that were restudy practiced and the right column items that were retrieval practiced in Session 1. Vertical dotted lines represent recall threshold; items that are above threshold are recalled, items below threshold are not. The top pair of panels depicts memory strength distributions after initial study in Session 1; at this point, the two distributions are identical. The second pair of panels shows how distributions are shifted after restudy or retrieval practice. Restudied items are strengthened to the same degree, which moves the distribution to the right. In contrast, the distribution of retrieval-practiced items becomes bifurcated; successfully retrieved items are strengthened to a higher degree than restudied items, whereas not retrieved items remain at original memory-strength level. At the end of Session 1, more of the restudied than of the retrieval-practiced items are above recall threshold. The third pair of panels illustrates distributions after a relatively long retention interval (e.g., 2 days) in Session 2; all items are reduced in strength at a comparable rate, which moves the item distributions equally to the left. At this point, more of the retrieval-practiced than of the restudied items are above recall threshold, explaining the standard testing effect in Test 1 of Session 2. **b** Extension of the bifurcation model: Effects of feedback learning in Test 1. In the upper pair of panels, memory strength distributions after recall trials in Test 1 are illustrated. The distribution of restudied items becomes bifurcated; the successfully recalled items in both practice conditions are strengthened to a high degree. The bottom pair of panels shows how immediate feedback may affect items' memory strength. Critically, it is assumed that all items are strengthened to the same degree, regardless of items' memory strength and position relative to recall threshold

the testing effect, this study focusses on the more descriptive bifurcation model, partly because, unlike the semantic elaboration and episodic context accounts, it can account for both the testing effect and its reversal.

Possible effects of feedback learning arising from the bifurcation model

In explaining testing and reversed testing effects, the bifurcation model provides two suggestions on the effects of feedback learning in Test 1, leading to the reversal of the testing effect in Test 2. First, the bifurcation model suggests that it is the restudied items that are *not* recalled in Test 1 that generate the reversal of the testing effect in Test 2. This is because the not-recalled restudied items should be in close proximity to the recall threshold (see Fig. 1b), which makes it likely—more than for the relatively weaker retrieval-practiced items that are *not* recalled in Test 1—that feedback learning in Test 1 moves them beyond the recall threshold and thus makes them recallable in Test 2. In contrast, items that are successfully recalled in Test 1 should not contribute to the reversed testing effect, because these items—in both practice condition—already are above the recall threshold and thus perfectly recallable in Tests 1 and 2, regardless of whether or not feedback learning enhances their memory strength even further. Importantly, whether the reversed testing effect in Test 2 is indeed selectively reflected in a restudy over retrieval-practice benefit for those items that are *not* recalled in Test 1 has not yet been examined. A first aim of this study was thus to close this gap by running conditional analyses on the effects of practice condition on recall rates in Test 2 separately for items that were recalled and items that were not recalled in Test 1.

Second, the bifurcation model suggests that feedback learning in Test 1 enhances the memory strength of both recalled and not-recalled items (see Fig. 1b). Indeed, according to the model, a restudy opportunity may strengthen all items equally, moving the distribution(s) of items (equally) to the right on the memory-strength dimension (Kornell et al., 2011).[1] A restudy opportunity should therefore increase items' memory strength largely independent of the items' original memory-strength level and the items' position relative to the recall threshold. Translated to the reversed testing effect (Storm et al., 2014), this means that a restudy opportunity, provided in the form of feedback learning in Test 1, would strengthen all initially studied items equally, independent of whether the items have been repeatedly retrieved (and thus are relatively weak or relatively strong, depending on items' retrieval success during practice) or restudied (and thus are more moderate in memory strength) in Session 1, and also independent of whether the items have been successfully recalled (and thus are above the recall threshold) or not (and thus are below the recall threshold) in Test 1 of Session 2. Unfortunately, these predictions cannot be tested by examining behavioral recall accuracy data alone. In fact, while recall accuracy may be able to differentiate between items that are above and items that are below the recall threshold, it does not measure memory strength per se and thus may not be able to indicate whether or not feedback strengthens single item types equally. To test the model's predictions, however, neurocognitive methods can be used. A second aim of the present study thus was to test the bifurcation model's predictions by investigating EEG brain oscillatory activities related to feedback learning and the reversed testing effect.

Neurocognitive studies

While a number of recent studies examined the neural underpinnings of the testing effect, by investigating both the BOLD signal in functional magnetic resonance imaging (fMRI; Eriksson, Kalpouzos, & Nyberg, 2011; Hashimoto, Usui, Taira, & Kojima, 2011; Keresztes, Kaiser, Kovács, & Racsmány, 2014; van den Broek, Takashima, Segers, Fernández, & Verhoeven, 2013; Wimber, Rutschmann, Greenlee, & Bäuml, 2009; Wing, Marsh, & Cabeza, 2013; Wirebring et al., 2015) and event-related potentials in the electroencephalogram (EEG; Bai, Bridger, Zimmer, & Mecklinger, 2015; Rosburg, Johansson, Weigl, & Mecklinger, 2015), there is no neurocognitive study that examined the reversal of the testing effect to date. In the present study, we examined the neural signature of the reversed testing effect by investigating EEG brain oscillations. Brain oscillations play a major role in the synchronization of neural networks, and they are crucially involved in the formation, consolidation, and retrieval of memories (Fries, 2005; Fell & Axmacher, 2011). Thereby, brain oscillations at different frequencies have been shown to be related to episodic memory, both at encoding and retrieval (for reviews, see Hanslmayr & Staudigl, 2014; Nyhus & Curran, 2010). For instance, at encoding, power decreases of cortical EEG alpha (~10 Hz), and lower-beta oscillations (~15 Hz) during item encoding have been linked to semantic elaboration and deep encoding of item information (Hanslmayr, Spitzer, & Bäuml, 2009; Klimesch, Doppelmayr, Pachinger, & Russegger, 1997), with lower EEG beta power decreases during item encoding being associated with increases in BOLD activity in the left ventrolateral prefrontal cortex (Hanslmayr et al., 2009), which is a brain region that has been closely linked to elaborative processing of item information (Han, O'Connor, Eslick, & Dobbins, 2012; Otten & Rugg, 2001). In addition to alpha/lower-beta oscillations, stimulus-induced power increases of EEG theta oscillations (3–7 Hz) have been linked to memory encoding. Functionally, these increases of theta power during item encoding have been linked to the binding of item-to-context information (Summerfield & Mangels, 2005; Staudigl & Hanslmayr, 2013).

In addition to encoding, brain oscillations have been shown to play an important role for retrieval. In particular, brain oscillations in the traditional theta frequency range (3–7 Hz) have been linked to episodic memory retrieval. Thereby, both positive and negative relationships between theta power increases during item retrieval and retrieval success have been reported in previous work, with positive effects in theta power being functionally linked to processes of recollection and conscious awareness (Gruber, Tsivilis, Giabbiconi, & Müller, 2008; Klimesch et al., 2001) and negative effects to processes of interference and cognitive control (Hanslmayr, Staudigl, Aslan, & Bäuml, 2010; Staudigl, Hanslmayr, & Bäuml,

[1] The bifurcation model has been introduced in the two recent research papers by Kornell et al. (2011) and Halamish and Bjork (2011). The suggestion that a restudy opportunity strengthens all items equally is taken from Kornell et al. (2011), who assumed that "the restudy items . . . all gain memory strength equally", whereas "the tested items . . . become bifurcated" and the retrieved items gain more strength than the restudied items (see caption of Figure 1 in Kornell et al., 2011, p. 87). Halamish and Bjork (2011) were silent on the issue and did not assume that all restudied items are strengthened equally. When we refer to the idea that a restudy opportunity (or feedback learning) strengthens all items equally, we thus refer to the bifurcation model's assumptions as introduced by Kornell et al. (2011).

2010). Reconciling these findings, Pastötter and Bäuml (2014) recently showed that distinct slow and fast theta oscillations—at the edges of the traditional theta frequency band—are differentially related to retrieval success (see also Lega, Jacobs, & Kahana, 2012). Increases of slow theta power (~3 Hz) show a positive relationship, whereas increases of fast theta power (~7 Hz) show a negative relationship, suggesting that slow and fast theta oscillations have distinct functional roles in human episodic memory retrieval, with slow theta oscillations being related to recollection and conscious awareness and fast theta oscillations being related to conflict monitoring and cognitive control.

The present study

In this study, we examined the EEG oscillatory correlates of the reversed testing effect (Storm et al., 2014) and tested the distribution-based bifurcation model's explanation of testing and reversed testing effects (Kornell et al., 2011). Closely following the procedure by Storm et al. (2014), participants took part in two experimental sessions. In Session 1, participants studied weakly related (German) word pairs (e.g., linen–towel) and then underwent a retrieval/restudy practice phase, in which half of the items were retrieval practiced and half were restudy practiced. Two days later, in Session 2, all of the intially studied items were tested in two final cued recall tests (Tests 1 and 2). Immediate feedback was provided to participants in Test 1, such that each recall trial (e.g., linen–____) was followed by a feedback trial (e.g., linen–TOWEL), regardless of whether or not an item was recalled. Scalp EEG was measured in Session 2. EEG data were analyzed in the time-frequency domain, with focus on stimulus-induced power changes in the slow theta, fast theta, and alpha/beta frequency ranges. Power changes were related to feedback learning in Test 1 and retrieval success in Tests 1 and 2, and were examined as a function of practice condition to investigate the neural signature of the reversed testing effect.

Several expectations arose with regard to behavioral recall data. First, we expected to replicate the findings by Storm et al. (2014), that is, a testing effect in Test 1 but a reversed testing effect in Test 2. Second, we expected the reversed testing effect in Test 2 to arise from a restudy over retrieval-practice benefit for items that were *not* recalled in Test 1. No such benefit was expected for items that were recalled in Test 1. Such a finding would be consistent with the bifurcation model's explanation of the reversed testing effect, according to which it is only the items that were not recalled in Test 1 whose recall should benefit from feedback, with the restudied items' recall showing a larger benefit from feedback than that of the retrieval-practiced items. In fact, the restudied items are in close proximity to the recall threshold, which makes it likely that feedback strengthens them beyond threshold and makes them recallable in Test 2. We conducted conditional analyses to examine the effects of retrieval versus restudy practice on recall accuracy in Test 2 separately for items that were recalled and items that were not recalled in Test 1.

Expectations arose also with regard to EEG data. First, following the semantic elaboration view on testing and test-potentiated learning effects (Carpenter, 2009; McDaniel & Masson, 1985) and the finding that decreases of alpha/lower-beta power are linked to the strengthening of items through semantic elaboration (Hanslmayr et al., 2009; Klimesch et al., 1997), an effect in alpha/lower-beta power during feedback learning in Test 1 was expected. Second, following the recent dissociation on retrieval-related theta effects (Pastötter & Bäuml, 2014), concurrent effects in slow and fast theta power were expected, with slow theta power being positively and fast theta power being negatively related to retrieval success. Third, and most important, following the bifurcation model's suggestion that feedback learning in Test 1 should increase memory strength of all items approximately equally, that is, independent of the items' original memory-strength level and the items' position relative to the recall threshold (Kornell et al., 2011), we expected to find the feedback-related effect in alpha/lower-beta power to be present regardless of practice condition and of whether items were recalled or not. For this aim, conditional EEG analyses were calculated and the effects of practice condition on feedback-related alpha/lower-beta power were examined separately for items that were recalled and items that were not recalled in Test 1. Finally, conditional EEG analyses were also calculated for retrieval-related effects in slow and fast theta power. According to the bifurcation model, the reversed testing effect in Test 2 arises from a restudy over retrieval-practice benefit only for those items that were *not* recalled in Test 1. Thus, we expected to find any effects of practice condition on retrieval-related theta power modulations to be restricted to items that were not recalled in Test 1.

Method

Participants

Twenty-six students (18 females) at Regensburg University, Germany, participated in this study. Participants' mean age was 23.5 years (*SD* = 2.5), ranging from 19 to 29 years. All participants reported normal or corrected-to-normal vision and spoke German as their native language. No participant reported any history of neurological disease. Twenty-two participants were right-handed, and four participants were left-handed. Participants gave written informed consent and were paid 25 Euros for participation. The study was conducted in accordance with the Declaration of Helsinki.

Stimuli

The stimuli were 120 (German translations of) weakly related cue–target word pairs taken from the Nelson, McEvoy, and Schreiber (2004) database (e.g., *linen–towel*, *mouse–hole*). The forward association strength of the word pairs was within a range of .041 to .058, meaning that when presented with the cue word, approximately 5 % of participants produced the target word as their first free associate in the Nelson et al. study. The assignment of word pairs to practice conditions was randomized across participants. Stimuli were presented with E-Prime software (v2.0.10.353, Psychology Software Tools, Sharpsburg, Pennsylvania, USA).

Procedure

The experiment consisted of two sessions, with a delay of 48 hours between them. Sessions took place in the same EEG laboratory. In both sessions, participants were seated in front of a 15-inch computer screen at a distance of 1.25 m. A computer keyboard was placed within reach. Scalp EEG was recorded in Session 2.

Session 1

Session 1 consisted of an initial study phase and a retrieval/restudy practice phase (see Fig. 2). In the initial study phase, participants studied 120 weakly related word pairs (e.g., *linen*–TOWEL*, mouse*–HOLE). Participants were told that they would later be tested on their ability to remember the right words, that is, the targets (e.g., TOWEL), when given the left words as cues (e.g., *linen*). Word pairs were presented one at a time in the center of the screen for 7 s each. Stimulus presentation was preceded by a 0.5-s prestimulus interval showing a fixation cross in the middle of the screen. During the last 2 s of stimulus presentation, three response letter options (e.g., L, H, S) were shown at the bottom of the screen, and participants were instructed to indicate a target's last letter from the three letter options by pressing the corresponding key on the numeric keypad with the right-hand index, middle, and ring fingers. The letter-indication task was included in the initial study phase to familiarize participants with this response procedure, which was also subsequently used during recall testing.

In the retrieval/restudy practice phase, participants were given two blocks of randomly intermixed retrieval and restudy practice trials. In each block, 60 word pairs with the target missing, that is, only the cues, were shown for retrieval practice (e.g., *linen*–____), and 60 intact word pairs were shown for restudy practice (e.g., *mouse*–HOLE). The assignment of word pairs to practice conditions was randomized across participants but maintained across blocks. That is, each word pair was either restudied twice or retrieved twice. Word pairs with the target missing (i.e., cues only) and intact word pairs were shown one at a time in the center of the screen for 4.5 s each. Stimulus presentation was preceded by a 0.5-s prestimulus interval showing a fixation cross in the middle of the screen. During the last 2 s of stimulus presentation, three letter options were shown at the bottom of the screen, and participants were instructed to indicate the last letter of the target from the three letter options. Participants were to respond in both retrieval and restudy practice trials (i.e., both when the target was missing and when the target was shown). Participants were told not to guess. No feedback was provided. Responses were scored correct only if given within the 2 s in which response letters were shown.

Session 2 (with EEG)

Participants returned to the EEG lab 48 hours later. They were told that in each of two cued recall test (Tests 1 and 2) of Session 2 they would be tested on each of the 120 word pairs they had studied in the initial study phase of Session 1. Participants were informed that in Test 1, but not in Test 2, each test trial would be followed by a feedback trial in which the corresponding word pairs would be presented again, intact. Participants were also told that immediate feedback would be provided regardless of whether or not target recall in a preceding test trial was successful. Participants were asked to pay close attention to the encoding of items during feedback in Test 1 because all word pairs would be tested again in Test 2.

In Test 1, the 120 cues of the initially studied word pairs were shown one at a time in the center of the screen for 4 s each. Cue presentation was preceded by a 1.5-s prestimulus interval. Order of cues was randomized. During the last 2 s of cue presentation, three response letter options were shown, and participants indicated the last letter of the targets with their right-hand fingers. Participants were told not to guess. Immediate feedback was provided by showing the corresponding words pair intact for 2 s each. Feedback presentation was also preceded by a 1.5-s prestimulus interval. Responses were scored correct only if given within the 2 s in which response letters were shown.

In Test 2, all 120 cues were shown once again in a new randomized order. Cues were presented for 4 s each, separated by a 1.5-s prestimulus interval. As in Test 1, three letter options were shown during the last 2 s of cue presentation, and participants were to indicate the last letter of each target with their right-hand fingers. No feedback was provided in Test 2. Between Tests 1 and 2, participants did a simple 10-min choice reaction time task.

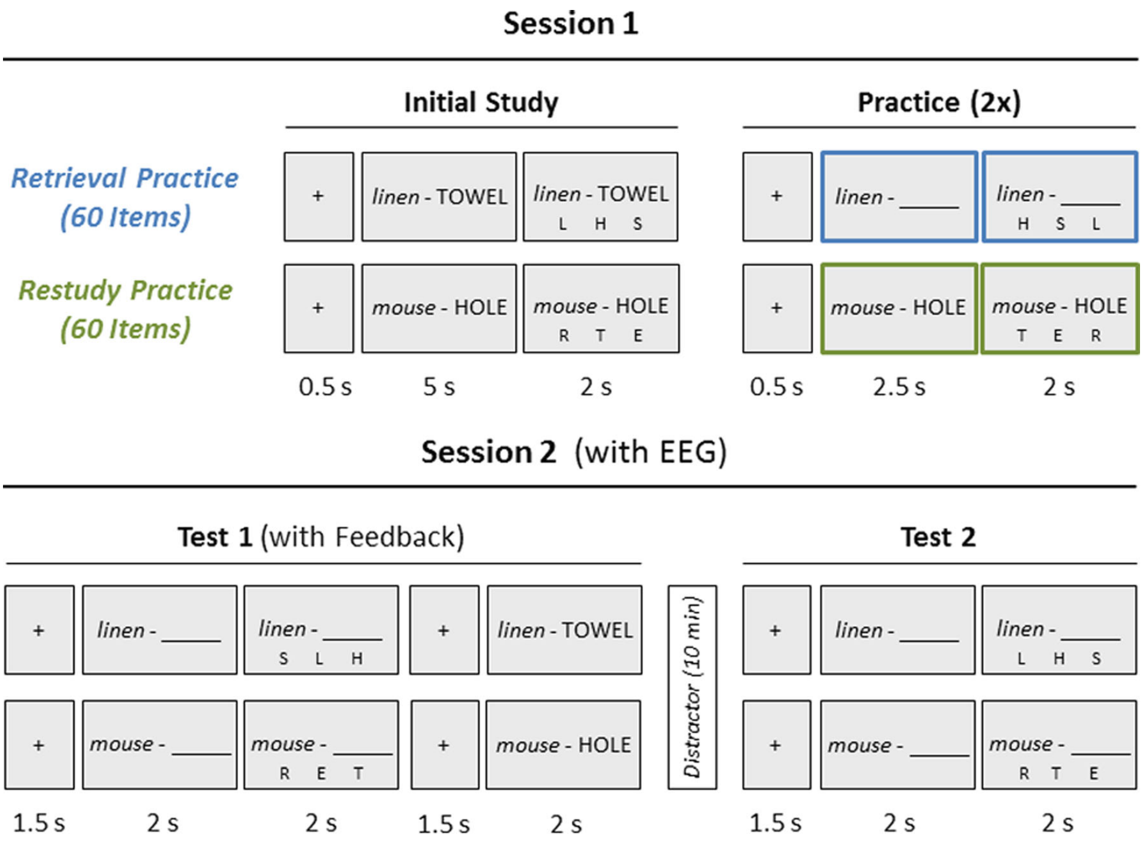

Fig. 2 Experimental procedure. In Session 1, participants studied 120 weakly related cue-target word pairs and then underwent a practice phase in which half of the word pairs were tested twice (retrieval-practice condition) and half were restudied twice (restudy-practice condition). Two days later, in Session 2, all initially studied items were tested twice in two cued-recall tests. Feedback was provided only in Test 1. In both sessions, participants were to type in responses by choosing the last letter of a target from three response letter options, regardless of whether the word pair was shown intact or with the target missing. Scalp EEG was measured in Session 2

Recording of EEG data

In Session 2, participants' scalp EEG was recorded from 61 equidistant active electrodes mounted in an elastic cap (ActiCAP, Montage 10, Brain Products, Gilching, Germany). ActiCAP is an active electrode system that enables fast electrode placement and low electrode–skin impedance due to amplification circuitry built into the electrodes. Electrode–skin impedance was kept below 20 kΩ. Electrode Cz served as common reference. Signals were digitized with a sampling rate of 500 Hz and amplified between 0.15 and 250 Hz with a notch filter at 50 Hz, removing power-line noise, which has a 50 Hz frequency in Europe (BrainAmpMR plus, BrainVision Recorder, v1.20, Brain Products, Gilching, Germany).

Preprocessing of EEG data

EEG recordings were rereferenced offline against average reference and were EOG corrected using calibration data and individual EOG artifact coefficients (Ille, Berg, & Scherg, 2002), as implemented in the BESA Research software package (v6.0, BESA Software, Gräfelfing, Germany). Remaining artifacts in the EEG were marked by careful visual inspection. The EEG data were segmented into 5.3-s epochs ranging from -2 s before to 3.3 s following onset of stimuli (cues, word pairs). However, in order to avoid filter artifacts at the edges of segments, all further analyses were restricted to intervals from -0.7 s to 2 s around stimulus onset. Segments containing intervals with marked artifacts were excluded from further analyses.

Both overall and conditional EEG analyses were calculated (see Statistical analysis of EEG data below). In the overall feedback-learning analysis, mean number of analyzed segments per subject was 45.5 (ranging from 21 to 78) for the previously recalled items and 63.3 (ranging from 21 to 98) for the previously not-recalled items (combined across practice conditions in Test 1). In the overall retrieval-success analysis, mean number of segments per subject was 121.7 (ranging from 72 to 187) for recalled items and 93.5 (ranging from 42 to 156) for the not-recalled items (combined across practice conditions and recall tests). Overall analyses were based on EEG data from all 26 subjects. In conditional feedback-learning and retrieval-success analyses, EEG data were examined separately for items that were (previously) recalled and items that were (previously) not recalled in Test 1. Each conditional analysis was restricted to participants who provided at least 20 artifact-free segments per calculation. Due to this restriction, four participants were excluded from conditional analyses of items that were (previously) recalled in Test 1, and four different participants were excluded from conditional analyses of items that were (previously) not recalled in Test 1. To rule out systematic variance in the number of trials between experimental conditions, in all conditional EEG analyses, number of trials in each condition was equated by selecting a subsample of trials that was equal to the number of trials in the condition with the fewest trials.

Time-frequency decomposition

EEG data were transformed into the time-frequency domain using a complex demodulation algorithm implemented in BESA Research (v6.0.04.2014; Hoechstetter et al., 2004). The algorithm consists of a multiplication of the time-domain signal with a complex periodic exponential function, having a frequency equal to the frequency under analysis, and subsequent low-pass filtering. The low-pass filter is a finite impulse response filter of Gaussian shape in the time domain, which is related to the envelope of the moving window in wavelet analysis. The data were sampled in the frequency range from 1 to 20 Hz and exported in bins of 0.1 s and 0.5 Hz. Time resolution was set to 0.158 s (full power width at half maximum; FWHM) and frequency resolution was set to 0.708 Hz (FWHM).

Stimulus-induced power changes were determined by calculating temporal-spectral evolution, that is, power changes during stimulus presentation for all time-frequency points with power increases or decreases at time point *t* and frequency *f* related to mean power at frequency *f* over the prestimulus baseline interval (Pfurtscheller & Aranibar, 1977; Pfurtscheller & Lopes da Silva, 1999). The baseline interval was set from -0.7 s to -0.2 s before stimulus onset, both for the presentation of word pairs and the presentation of cues.

Statistical analysis of EEG data

To control for problems of multiple comparisons when testing the significance of power differences over multiple time-frequency points and electrode sites, cluster and random permutation analyses were conducted (Maris & Oostenveld, 2007), using the software package BESA Statistics (v1.0, BESA Software, Gräfelfing, Germany).

Three steps were taken in the statistical analysis. First, nonspatial overall cluster analyses were calculated, in which time-frequency data were averaged across all 61 electrodes and contrasted between item types ([previously] recalled items, [previously] not-recalled items). In the overall feedback-learning analysis, time-frequency spectrograms of power changes during the presentation of intact word pairs were compared between previously recalled and previously not-recalled items (in Test 1). In the overall retrieval-success analysis, time-frequency spectrograms of power changes during cue presentation were compared between recalled and not-recalled items (combined across Tests 1 and 2). In both overall analyses, *t* tests were calculated for each time-frequency point (1,404 = 36 [time bins] * 39 [frequency bins]). For each cluster analysis of significant time-frequency windows, only adjacent time-frequency points that fell below a *p* value of .01 in the *t* test were considered. For each cluster, the sum of *t* values of the single significant time-frequency points was calculated as a test statistic. In random permutation analysis, 5,000 random permutations were run in which this statistic was repeated for randomly shuffled data sets in which data were randomly reordered across item types ([previously] recalled items, [previously] not-recalled items) and the cluster with the highest sum of *t* values was kept. By these means, null distributions were created from the 5,000 random permutation runs, and the critical p_{rand} values for the empirically derived time-frequency clusters were calculated.

In the next step, spatial cluster analyses were calculated to examine topographies of significant time-frequency clusters from the overall analyses. For each significant time-frequency window, power changes were averaged across time-frequency points, and averaged power changes were contrasted between item types ([previously] recalled items, [previously] not-recalled items) by calculating *t* tests for each electrode site (61). Spatial clusters were identified by considering only those contiguous electrode sites (with maximum distance of 45 mm between neighboring sites, resulting in an average of 4.87 neighbors per electrode site) that fell below a *p* value of .05 in the *t* test. For each spatial cluster, the sum of *t* values of the single contiguous electrode sites was calculated as a test statistic. Calculation of p_{rand} was based on 5,000 random permutation runs.

In the third step, conditional analyses were calculated in which time-frequency data were examined separately for items that were (previously) recalled and items that were (previously) not recalled in Test 1. For each spatial cluster from the analyses, power changes were averaged across significant electrode sites. In conditional feedback-learning analyses, averaged power changes were analyzed as a function of CONDITION (retrieval practice, restudy practice). In conditional retrieval-success analyses, averaged power changes were examined as a function of CONDITION (retrieval practice, restudy practice) and TEST (Test 1, Test 2).

Results

Behavioral results

In Session 1, participants correctly recalled 45.8 % (*SE* = 4.0 %) of the retrieval-practiced items in the first block and 51.2 %

(*SE* = 3.9 %) in the second block of the retrieval/restudy-practice phase (see left panel of Fig. 3). The increase in recall across blocks was significant (t_{25} = 4.35, *p* < .001, *d* = .85).

In Test 1 of Session 2, participants correctly recalled 46.4 % (*SE* = 3.7 %) of the retrieval-practiced items and 36.5 % (*SE* = 3.3 %) of the restudied items, indicating a significant testing effect, *t*(25) = 4.74, *p* < .001, *d* = .93 (see middle panel of Fig. 3). In Test 2, participants correctly recalled 68.3 % (*SE* = 3.2 %) of the retrieval-practiced items and 73.9 % (*SE* = 2.6 %) of the restudied items, indicating a significant reversal of the testing effect, *t*(25) = 3.65, *p* = .001, *d* = .71. When calculating a repeated-measures analysis of variance (ANOVA) with the factors of CONDITION (retrieval practice vs. restudy practice) and TEST (Test 1 vs. Test 2), the reversal of the testing effect was reflected in a significant crossover interaction between factors, *F*(1, 25) = 69.02, *p* < .001, partial η^2 = .73. The ANOVA also showed a main effect of TEST, *F*(1, 25) = 286.63, *p* < .001, partial η^2 = .92, which indicates that feedback learning in Test 1 enhanced recall from Test 1 to Test 2 but no significant main effect of CONDITION, *F*(1, 25) = 1.86, *p* = .185.

Conditional recall analyses further showed that the reversal in Test 2 arose from a restudy over retrieval-practice benefit for those items that were not recalled in Test 1 (restudy practice: *M* = 64.7 %, *SE* = 2.8 %; retrieval practice: *M* = 51.0 %, *SE* = 3.5 %), *t*(25) = 5.84, *p* < .001, *d* = 1.15. No such effect of practice condition in Test 2 arose for items that were successfully recalled in Test 1 (restudy practice: *M* = 91.3 %, *SE* = 2.0 %; retrieval practice: *M* = 91.2 %, *SE* = 1.4 %), *t*(25) < 1 (see right panel of Fig. 3).

Electrophysiological results

Overall analyses

Nonspatial cluster analyses were calculated to identify time-frequency windows of significant effects in brain oscillatory activity related to feedback learning and retrieval success, with time-frequency data averaged across all 61 electrodes. In the overall feedback-learning analysis, power changes following stimulus onset in feedback trials in Test 1 were contrasted between items that were recalled and items that were not recalled in preceding test trials. The analysis revealed two positive-going effects with higher power values during feedback of previously recalled than previously not-recalled items: a first effect in the slow theta frequency range (2 Hz to 4 Hz, time range: 0.2 s to 0.8 s), and a second effect in the alpha/lower-beta frequency range (12 Hz to 16 Hz, time range: 0.5 s to 1 s; see left panel of Fig. 4). In the overall retrieval-success analysis, power changes following cue onset were compared between recalled and not-recalled item, combined across Tests 1 and 2. The analysis revealed a positive effect in the slow theta frequency range (2 Hz to 4 Hz, time range: 0.6 s to 0.9 s), with higher power values for recalled than for not-recalled items, and two negative effects with lower power values for recalled than for not-recalled items: one in the fast theta frequency range (4.5 Hz to 8.5 Hz, time range: 1 s to 2 s) and the other in the beta frequency range (14 Hz to 18 Hz, time range: 0.6 s to 2 s; see right panel of Fig. 4). All further analyses were based on time-frequency windows of these significant overall effects.

Feedback learning

Spatial cluster analyses were calculated to examine topographies of overall feedback-learning effects. Figure 5 (left panel) visualizes topographies of the effects in slow theta and alpha/lower-beta power. Spatial analyses revealed a mid-frontocentral cluster of electrodes showing a larger increase of stimulus-induced slow theta power during feedback presentation of previously recalled items than of previously not-recalled items (p_{rand} < .005), and a tempo-parietal cluster of electrodes showing a larger decrease of alpha/lower-beta power during feedback presentation of the previously *not* recalled than the previously recalled items (p_{rand} < .001).

Conditional EEG analyses were calculated to examine whether feedback effects in slow theta and alpha/lower-beta were modulated by practice condition separately for items that were successfully recalled and items that were not recalled in Test 1. For each item type, power changes were averaged across significant electrodes of spatial clusters and examined as a function of practice condition. Results showed that neither slow theta nor alpha/lower-beta power differed between practice conditions, both for the recalled and the not-recalled items, all *t*(21)s < 1.3 (see middle and right panels of Fig. 5). While the previously *not*-recalled items showed a reliable decrease in stimulus-induced alpha/lower-beta power (*M* = -19.9 %, *SE* = 3.8 %), *t*(21) = 5.24, *p* < .001, *d* = 1.12, there was no significant change in stimulus-induced alpha/lower-beta power during feedback presentation of the previously recalled items (*M* = -3.6 %, *SE* = 4.8 %), *t*(21) = 0.75, *p* = .459.

Retrieval success

Figure 6 (left panel) shows topographies of the overall retrieval effects, collapsed across Tests 1 and 2. Spatial analyses revealed a mid-frontocentral cluster of electrodes for the positive effect in slow theta power (p_{rand} < .01), a left-to-mid-frontal cluster for the negative effect in fast theta power (p_{rand} < .001), and a left-to-mid-central cluster for the negative effect in beta power (p_{rand} < .001). Note that because, in the fast theta and beta power analyses, all or almost all of the electrodes fell below the critical *p* value of .05 in the single *t* test calculation, a stricter criterion threshold was used and the critical *p* value was set to .001 for single-electrode

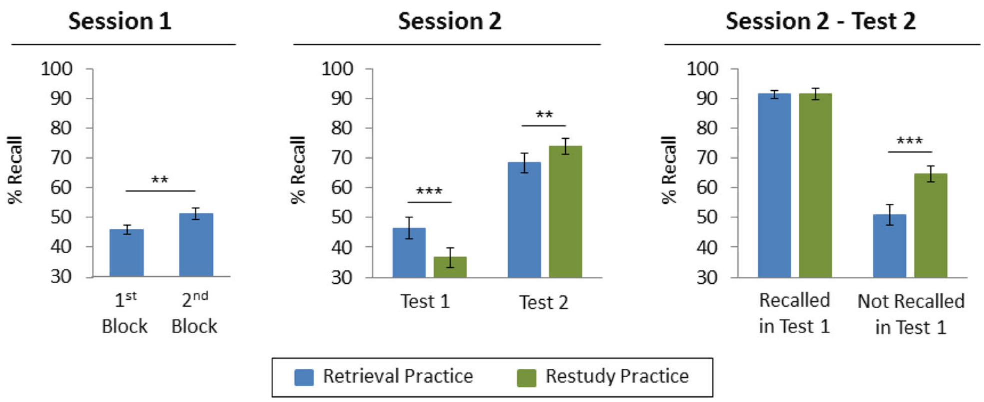

Fig. 3 Behavioral recall results. Left panel: In Session 1, participants correctly recalled about 50 % of the retrieval-practiced items. Recall increased across blocks. Middle panel: In Session 2, a reliable testing effect arose in Test 1, but a reversed testing effect arose in Test 2. Right panel: Conditional analyses of recall in Test 2 showed that the reversed testing effect arose from a practice effect for items that were *not* recalled in Test 1; no such practice effect arose for items that were successfully recalled in Test 1. Error bars represent the standard error of the mean. ** $p < .01$. *** $p < .001$

comparisons in these spatial analyses (depicted as white electrode clusters in Fig. 6; please note that all conditional effects, as reported subsequently, replicated when the .05 criterion threshold was used in fast theta and beta power analyses).

Conditional EEG analyses were calculated to examine whether retrieval effects in slow theta, fast theta, and beta power were modulated by practice condition separately for items that were recalled and items that were not recalled in Test 1. Power changes were averaged across significant electrodes of each spatial cluster. Six separate ANOVAs were calculated in which averaged power data in the three time-frequency ranges were examined as a function of practice CONDITION (retrieval practice vs. restudy practice) and TEST (Test 1 vs. Test 2) separately for items that were recalled and items that were not recalled in Test 1. Results showed no main effects or interactions for items that were recalled in Test 1, all Fs(1, 21) < 1.4 (see middle panel in Fig. 6). However, significant main effects and also a significant interaction arose for items that were *not* recalled in Test 1. For those *not*-recalled items, ANOVAs showed significant main effects of TEST, due to increases of slow theta power. $F(1, 21) = 6.40$, $p = .019$, partial $\eta^2 = .23$, and decreases of fast theta, $F(1, 21) = 24.22$, $p < .001$, partial $\eta^2 = .54$, and beta power across tests, $F(1, 21) = 26.08$, $p < .001$, partial $\eta^2 = .55$. In addition, the analyses showed a main effect of CONDITION for fast theta power, such that fast theta power was higher for retrieval-practiced items than for restudied items, $F(1, 21) = 4.53$, $p = .045$, partial $\eta^2 = .18$. Third, and most importantly, the main effect in fast theta power was qualified by a significant interaction between the factors of CONDITION and TEST, $F(1, 21) = 6.32$, $p = .020$, partial $\eta^2 = .23$, indicating a larger decrease in fast theta power across recall tests for the restudied than for the retrieval-practiced items. Post hoc contrasts further showed that there was a reliable decrease in fast theta power across tests for both the restudied items (Test 1: $M = 22.7$ %, $SE = 5.1$ %; Test 2: $M = 0.6$ %, $SE = 4.3$ %), $t(21) = 4.21$, $p < .001$, $d = .90$, and the retrieval-practiced items (Test 1: $M = 21.6$ %, $SE = 3.2$ %; Test 2: $M = 15.3$ %, $SE = 3.6$ %), $t(21) = 2.14$, $p = .045$, $d = .46$. Fast theta power differed between conditions in Test 2, $t(21) = 3.56$, $p = .002$, $d = .76$), but not in Test 1, $t(21) < 1$. ANOVAs showed no other significant main effects or interactions, all Fs(1, 21) < 1.[2]

Discussion

Implications from behavioral results

Behavioral recall results replicate the recent findings by Storm et al. (2014), demonstrating a significant testing effect (i.e., better recall of retrieval-practiced than restudied items) in Test 1 but a reversed testing effect (i.e., better recall of restudied than retrieval-practiced items) in Test 2. To our knowledge, this is the first replication of the Storm et al. finding, showing that the testing effect can be dramatically reversed when feedback is provided to participants and a second recall test is conducted during final recall testing. Moreover, because, relative to the study by Storm et al. (2014), the

[2] Following a reviewer's suggestion concerning the standard testing effect, we also analyzed Test 1 differences in retrieval-related slow theta, fast theta, and beta power as a function of practice condition, irrespective of item retrieval success. Data included in these analyses were based on the same electrodes and time-frequency clusters as were used in the corresponding conditional EEG analyses. All subjects were included. The results showed no significant effects. Still, there were tendencies for testing effects that were consistent in direction with the present retrieval-related reversed-testing effects. Specifically, retrieval-practiced items compared to restudied items showed a nonsignificant increase of slow theta power (27.7 % vs. 23.6 %), $t(25) = 1.15$, $p = .261$, and nonsignificant decreases of fast theta power (9.8 % vs. 16.6 %), $t(25) = 1.05$, $p = .304$, and beta power (-14.5 % vs. -11.9 %), $t(25) = 1.44$, $p = .161$.

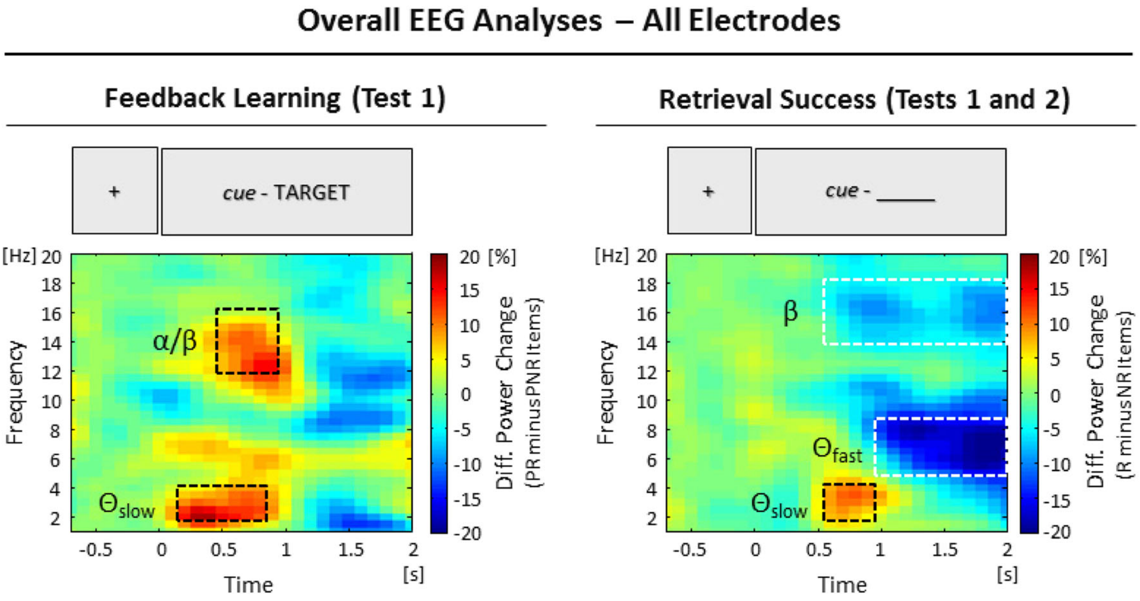

Fig. 4 Results from the overall EEG analyses with time-frequency spectrograms of power differences, time-locked to stimulus onset, *averaged over all electrodes*. Feedback analysis revealed two significant clusters with higher power values for previously (in Test 1) recalled (PR) than for previously not-recalled (PNR) items: one in the slow theta frequency range (2–4 Hz), and a second in the alpha/lower-beta frequency range (12–16 Hz). Retrieval-success analysis revealed one positive cluster in the slow theta frequency range (2–4 Hz), with higher power values for recalled (R) than for not-recalled (NR) items, and two negative clusters in the fast theta (4.5–8.5 Hz) and the beta frequency ranges (14–18 Hz), with lower power values for recalled (R) than for not-recalled (NR) items, combined for Tests 1 and 2

reversal was found with different materials, fewer retrieval/restudy practice trials and higher retrieval success rates in Session 1, and a shorter retention interval, the present results suggest generalizability of the reversed testing effect across different materials, task difficulties, and delays between study and final test. Going beyond the findings reported by Storm et al. (2014), results from the conditional recall analyses further showed that the reversal of the testing effect in Test 2 arose from a restudy over retrieval-practice benefit only for those items that were *not* recalled in Test 1. In fact, in Test 2, no effect of practice condition arose for items that were successfully recalled in Test 1.

The behavioral results are consistent with the distribution-based bifurcation model (Halamish & Bjork, 2011; Kornell et al., 2011). According to the model, the testing effect in Test 1 arises because more of the restudied than the successfully retrieval-practiced items fall below the recall threshold after delay and thus less of the restudied than the retrieval-practiced items are recallable in Test 1, which is what the present recall results show. With respect to the reversal of the testing effect in Test 2, the model assumes that more of the restudied items than the unsuccessfully retrieval-practiced items move over the recall threshold after feedback, and thus more of the restudied than the retrieval-practiced items become recallable in Test 2, which also is what the present results show. In particular, the model assumes that the restudied items that are *not* recalled in Test 1 generate the reversal of the testing effect in Test 2. This is because the *not*-recalled restudied items—in contrast to the relatively weaker not-recalled retrieval-practiced items—are in close proximity to the recall threshold before feedback is provided in Test 1 (see Fig. 1b), which makes it likely that feedback moves them over the recall threshold and thus makes them recallable in Test 2.[3] In contrast, the items that are successfully recalled in Test 1 should not contribute to the reversal. The present results from the conditional recall analyses are totally consistent with these assumptions and thus support the distribution-based bifurcation model of testing effects.

Implications from EEG results: feedback learning

To our knowledge, this is the first study to examine feedback-related brain oscillatory activity in an episodic memory task. Two significant effects arose with respect to differences in feedback-related brain activity between recalled and not-recalled items in Test 1. The first effect arose from a stimulus-induced decrease of alpha/lower-beta power (12 Hz to 16 Hz) that was present for the not-recalled items but absent for the recalled items. The second effect in feedback-related brain activity arose in the slow theta frequency range (2 Hz to 4 Hz), with the recalled items showing a larger increase than

[3] Still, the previously not-recalled retrieval-practiced items benefitted from feedback learning as well, with 51 % of them being recalled in Test 2, which qualifies the schematic representation of the bifurcation model as depicted in Figure 1B, in which none of the unsuccessfully retrieval-practiced items passed the recall threshold after feedback learning and thus none of them would become recallable in Test 2.

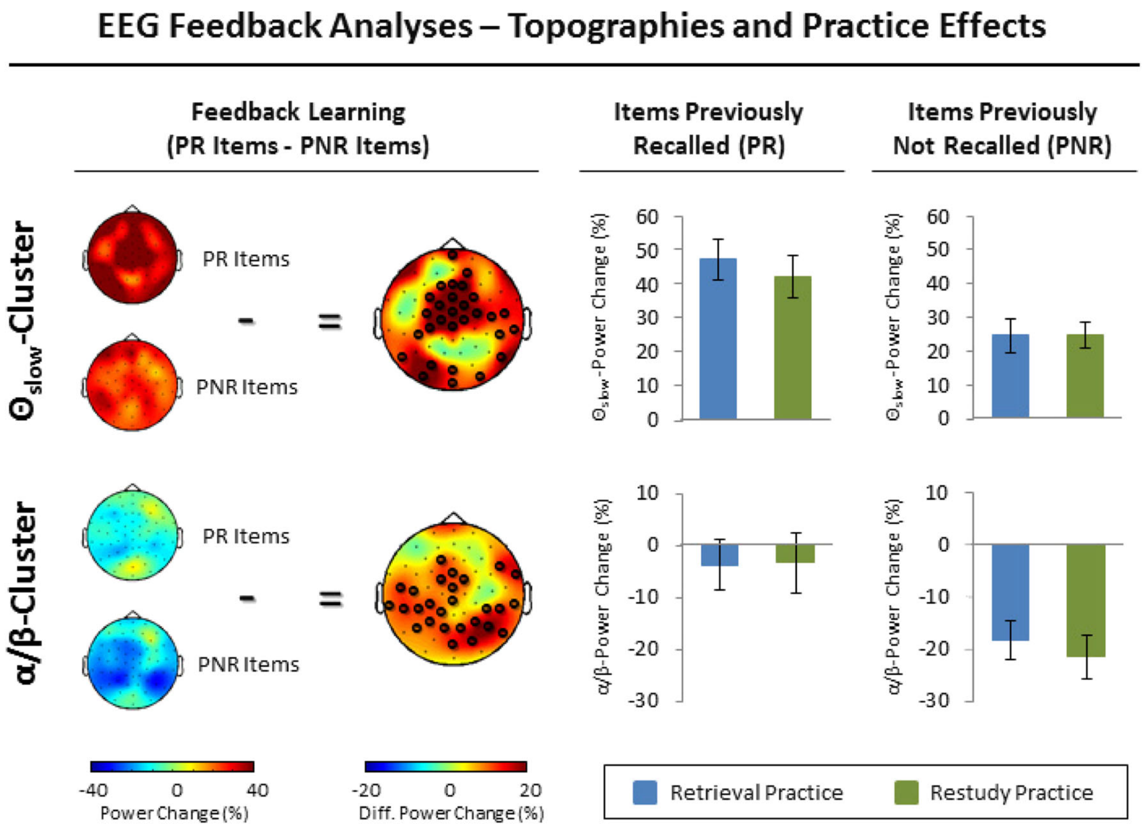

Fig. 5 Results from EEG feedback analyses. Left panel: Topographies and spatial clusters of feedback-learning effects in slow theta and alpha/lower-beta power, based on time-frequency ranges of significant effects from the overall feedback analysis. Right panel: Results from conditional EEG feedback analyses, showing stimulus-induced changes of slow theta and alpha/lower-beta power as a function of practice condition (retrieval practice, restudy practice) separately for PR and PNR items. Error bars represent the standard error of the mean

the not-recalled items. Both effects were unaffected by practice condition.

Because alpha/lower-beta power decreases have been related to semantic or deep item encoding in prior EEG work (Hanslmayr et al., 2009; Klimesch et al., 1997), the present alpha/lower-beta power effect is suggested to reflect strengthening of items via elaborative encoding processes, a view that is consistent with the semantic elaboration account of testing and test-potentiated learning effects (Carpenter, 2009; Pyc & Rawson, 2010). Importantly, the present alpha/lower-beta power results have theoretical implications for the distribution-based bifurcation model (Halamish & Bjork, 2011; Kornell et al., 2011). The bifurcation model assumes that a restudy opportunity strengthens all items equally (i.e., independent of the items' original memory strength and the items' position relative to the recall threshold). For the present task, the model thus predicts that a restudy opportunity, provided through feedback learning in Test 1, should strengthen all items equally (i.e., independent of whether the items have been retrieval practiced or restudy practiced in Session 1, and independent of whether the items have been recalled or not recalled in Test 1 of Session 2). While the present alpha/lower-beta results are consistent with the first prediction, they are inconsistent with the second. Indeed, the results showed no effect of practice condition on alpha/lower-beta power, indicating that feedback learning strengthened the retrieval-practiced and the restudied items to the same degree, which is consistent with the bifurcation model. However, the results further showed that only the items that were *not* recalled in Test 1, but not the items that *were* recalled in Test 1, showed a decrease in alpha/lower-beta power, indicating that feedback strengthens the *not*-recalled items, but not of the recalled items, which is inconsistent with the model. Together, the results suggest a modification of the bifurcation model, according to which a restudy opportunity—in the form of feedback learning in Session 2—strengthens all items below the recall threshold to the same degree and regardless of whether the items have been restudy practiced or retrieval practiced in Session 1, but leaves the items above recall threshold unaffected (see lower panel in Fig. 1b). Such modification of the model would also be consistent with findings from behavioral work showing that long-term benefits of feedback indeed can be strikingly absent once an item has been successfully recalled from memory (Karpicke & Roediger, 2008; Pashler, Cepeda, Wixted, & Rohrer, 2005).

Regarding the feedback-related effect in slow theta power, topography, direction, and size of the effect were highly similar to those observed for the slow theta effect related to

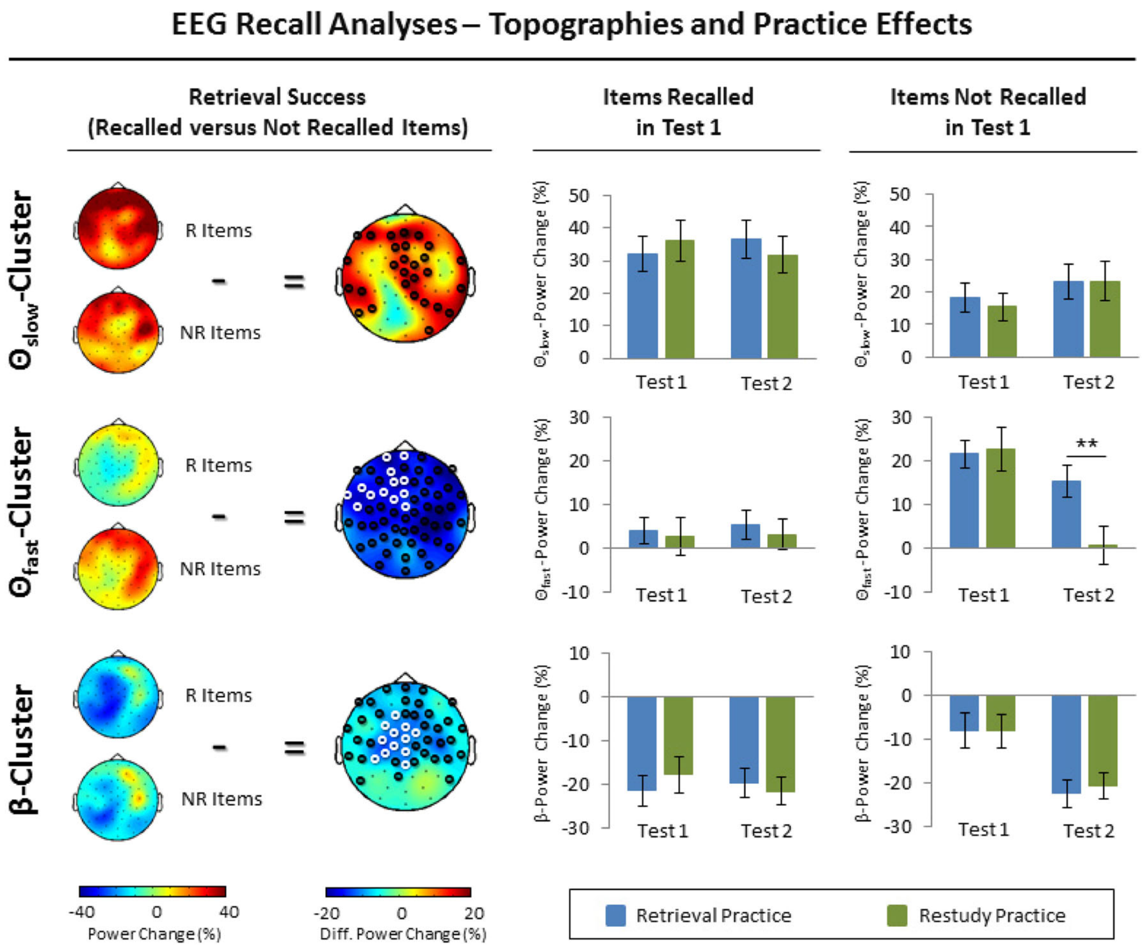

Fig. 6 Results from EEG retrieval-success analyses. Left panel: Topographies and spatial clusters of retrieval-success effects in slow theta, fast theta, and beta power, based on time-frequency ranges of significant effects from the overall retrieval-success analysis, combined for Tests 1 and 2. Right panel: Results from conditional EEG retrieval-success analyses, showing cue-induced changes of slow theta, fast theta, and beta power as a function of practice condition (retrieval practice, restudy practice) and test (Test 1, Test 2) separately for items that were recalled and items that were not recalled in Test 1. Error bars represent the standard error of the mean. ** $p < .01$

retrieval success. Therefore, we would suggest that the effect in slow theta power reflects retrieval-related processes (i.e., recollection) rather than processes related to encoding. In fact, a significant contribution of retrieval-related processes to feedback learning and test-potentiated learning has been suggested in previous imaging work (Nelson, Arnold, Gilmore, & McDermott, 2013). The exact relationship between theta brain oscillations and retrieval is discussed next.

Implications from EEG results: retrieval success

Previous EEG work has shown that theta oscillations during episodic memory retrieval are related to retrieval success. In particular, recent work revealed a dissociation between slow and fast theta oscillations, with increases of slow theta power (~3 Hz) being positively related to retrieval success, and increases of fast theta power (~7 Hz) being negatively related to retrieval success (Pastötter & Bäuml, 2014). The present EEG results replicate this dissociation in retrieval-related theta dynamics, demonstrating a larger stimulus-induced increase of slow theta power but a smaller stimulus-induced increase of fast theta power, for the recalled than for the not-recalled items. Indeed, although materials and retention interval differed between the present and our previous EEG study, topographies and time courses of slow and fast theta effects were very similar between studies. Theoretically, the present findings are consistent with the proposal that slow theta power reflects processes related to recollection and conscious awareness, whereas fast theta power reflects processes related to retrieval failure and cognitive control (Pastötter & Bäuml, 2014). In addition, results showed a retrieval-related effect in the beta frequency range (14 Hz to 18 Hz), dominant over left-to-mid-central electrode sites and characterized by a larger beta power drop for the recalled than for the not-recalled items. The effect was not influenced by practice condition. Given the left lateralization of the effect and because all responses were given with right-hand fingers, we suspect that the effect reflects processes related to response preparation. Indeed, it is a prominent finding that beta power decreases over (pre) motor sites are related to the preparation of motor responses (e.g., Pastötter, Berchtold, & Bäuml, 2012; Tzagarakis, Ince, Leuthold, & Pellizzer, 2010). Specifically, it has been shown that response preparation involves a decrease of beta activities that is typically more pronounced contralateral than ipsilateral to the responding hand, reflecting response activation or release of response inhibition (e.g., Doyle, Yarrow, & Brown, 2005; Stancák & Pfurtscheller, 1996).

The present fast theta results are consistent with the distribution-based bifurcation model (Halamish & Bjork, 2011; Kornell et al., 2011). In particular, conditional EEG analyses revealed that only the items that were *not* recalled in Test 1 but not the items that *were* recalled in Test 1 showed a modulation in fast theta power across recall tests, with the not recalled, restudied items showing a larger fast theta power decrease from Test 1 to Test 2 than the not recalled, retrieval-practiced items. Because increases of fast theta power index retrieval failure (Pastötter & Bäuml, 2014), the results indicate that it is the memory for the restudied items that were *not* recalled in Test 1 that benefits most from feedback learning in Test 1, with decreasing retrieval failure, or increasing retrieval success, for these items from Test 1 to Test 2. The modulation in fast theta power is perfectly consistent with the behavioral recall results, showing that it is the restudied items that are *not* recalled in Test 1 that show the largest increase in recall rates from Test 1 to Test 2, and it is also perfectly consistent with the bifurcation model, according to which the restudied items that are *not* recalled in Test 1 generate the reversal of the testing effect. Notably, from a processing view on the effect, the fast theta results are also consistent with the recently suggested mediator shift hypothesis of testing and test-potentiated learning effects (Pyc & Rawson, 2010, 2012), according to which participants monitor and evaluate the effectiveness of learning and modify feedback-related encoding strategy (e.g., from shallow to deep encoding as it is suggested by the semantic elaboration account) after retrieval failure. Indeed, increases of midfrontal (fast) theta power have been shown to index processes related to conflict monitoring and cognitive control in both memory (e.g., Hanslmayr et al., 2010; Pastötter & Bäuml, 2014; Staudigl, Hanslmayr, & Bäuml, 2010) and nonmemory tasks (e.g., Cohen & van Gaal, 2014; Pastötter, Dreisbach, & Bäuml, 2013; van Driel, Swart, Egner, Ridderinkhof, & Cohen, 2015).

With regard to the present effects in slow theta power, we follow the idea that retrieval-related increases of slow theta power index conscious recollection of memories (Pastötter & Bäuml, 2014). Thereby, slow theta power may reflect a qualitative index in terms of the presence or absence of conscious recollection rather than a quantitative index of items' memory strength. In fact, if slow cortical theta power reflected items' memory strength, an effect of practice condition for the successfully recalled items in Test 1, with higher power values for the retrieval-practiced than for the restudied items, should have been observed, at least if we follow the idea that the successfully retrieval-practiced items are strengthened to a higher degree than the restudied items (Halamish & Bjork, 2011; Kornell et al., 2011). Intracranial EEGs may be used to test the bifurcation model's assumptions concerning memory strength. Indeed, recent iEEG work has shown that theta activities in subcortical brain areas, including the hippocampus and the amygdala, are related more directly to items' memory strength (Rutishauser, Boss, Mamelak, & Schuman, 2010). Future work may thus examine the effects of retrieval versus restudy practice on items' memory strength by investigating iEEG. Finally, it should be noted that the present results seem to suggest that there was a systematic difference in retrieval-related effects in slow and fast theta power. Specifically, while increases of fast theta power were largely restricted to the not-recalled items, increases of slow theta power (and also beta power) were present for both recalled and not-recalled items. These results suggest that additional processes, which are not related to memory retrieval or recollection, may have contributed to the present slow theta effect. In fact, such contribution of additional, secondary processes may have increased error variance in the present slow theta analyses, which may explain the nonfinding of a reliable Test 2 practice effect for items that were not recalled in Test 1. Indeed, such slow theta effect, opposite in direction to the present fast theta effect, could have been expected. Future iEEG may help to disentangle the contribution of memory related and memory unrelated slow theta effects to testing and reversed testing effects.

Final remarks and conclusion

Both the present recall results and the EEG results are largely consistent with the distribution-based bifurcation model of testing and reversed testing effects (Halamish & Bjork, 2011; Kornell et al., 2011; Storm et al., 2014), though the alpha/lower-beta results imply some modification of the model with respect to the absence of an encoding-related feedback strengthening for items that have been successfully recalled in Test 1. According to this modification, only the not-recalled items, but not the recalled items, are strengthened by feedback, regardless of practice condition. With regard to the cognitive processes that may underlie the strengthening of items, and following the suggestion that alpha/lower-beta power decreases during item encoding index semantic or deep item encoding (Hanslmayr et al., 2009; Klimesch et al., 1997), the present results are particularly consistent with the semantic elaboration account of testing and test-potentiated learning effects (Carpenter, 2009; Pyc & Rawson, 2010), although other results clearly indicate an additional role of unique context cues at test, as it is suggested by the episodic context account of retrieval-based learning (e.g., Kliegl & Bäuml 2016;

Lehman et al., 2014). The bifurcation model is well compatible with the idea that both semantic elaboration and context updating via retrieval practice can contribute to memory strengthening of (the previously not-recalled) items.

Storm et al. (2014) argued that any use of certain standard ways to examine a research question can hide dynamics that are complex and interesting. In testing-effect studies, it is a standard way to assess participants' memory performance on a single, final criterion test. Moreover, it has been implied from these studies that recall performance on a single criterion test provides a good reflection of retrieval-based learning, especially when the test is given after a relatively long retention interval. Such implications, however, may not be warranted, as is suggested by the present EEG study and the recent behavioral study by Storm et al. (2014). Both studies consistently showed that the testing effect can be dramatically reversed when a restudy opportunity—in the form of feedback in a first criterion test—is provided and participants' memory is reassessed on a subsequent criterion test. Moreover, the present EEG study suggests that while (the neural markers of) feedback learning may not be affected by practice condition, memory performance and (the neural markers of) retrieval success can be. Memory performance thus may not (always) be a reliable measure of learning, and learning must be distinguished from performance and the retrievability of items (Bjork & Bjork, 1992; Soderstrom & Bjork, 2015).

References

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

Bai, C.-H., Bridger, E. K., Zimmer, H. D., & Mecklinger, A. (2015). The beneficial effect of testing: An event-related potential study. *Frontiers in Behavioral Neuroscience, 9,* 248.

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., & Bjork, E. L. (1992). A new theory of disuse and an old theory of stimulus fluctuation. In A. Healy, S. Kosslyn, & R. Shiffrin (Eds.), *From learning processes to cognitive processes: Essays in honor of William K. Estes* (Vol. 2, pp. 35–67). Hillsdale: Erlbaum.

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

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.

Carpenter, S. K., & Kelly, J. W. (2012). Tests enhance retention and transfer of spatial learning. *Psychonomic Bulletin & Review, 19,* 443–448.

Cohen, M. X., & van Gaal, S. (2014). Subthreshold muscle twitches dissociate oscillatory neural signatures of conflicts from errors. *NeuroImage, 86,* 503–513.

Doyle, L. M. F., Yarrow, K., & Brown, P. (2005). Lateralization of event-related beta desynchronization in the EEG during pre-cued reaction time tasks. *Clinical Neurophysiology, 116,* 1879–1888.

Dunlosky, J., Rawson, K. A., Marsh, E. J., Nathan, M. J., & Willingham, D. T. (2013). Improving students' learning with effective learning techniques: Promising directions from cognitive and educational psychology. *Psychological Science in the Public Interest, 14,* 4–58.

Eriksson, J., Kalpouzos, G., & Nyberg, L. (2011). Rewiring the brain with repeated retrieval: A parametric fMRI study of the testing effect. *Neuroscience Letters, 505,* 36–40.

Fell, J., & Axmacher, N. (2011). The role of phase synchronization in memory processes. *Nature Reviews Neuroscience, 12,* 105–118.

Fries, P. (2005). A mechanism for cognitive dynamics: Neuronal communication through neuronal coherence. *Trends in Cognitive Sciences, 9,* 474–480.

Gruber, T., Tsivilis, D., Giabbiconi, C. M., & Müller, M. M. (2008). Induced electroencephalogram oscillations during source memory: familiarity is reflected in the gamma band, recollection in the theta band. *Journal of Cognitive Neuroscience, 20*, 1043–1053.

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.

Han, S., O'Connor, A. R., Eslick, A. N., & Dobbins, I. G. (2012). The role of left ventrolateral prefrontal cortex during episodic decisions: Semantic elaboration or resolution of episodic interference? *Journal of Cognitive Neuroscience, 24,* 223–234.

Hanslmayr, S., Spitzer, B., & Bäuml, K.-H. (2009). Brain oscillations dissociate between semantic and non-semantic encoding of episodic memories. *Cerebral Cortex, 19,* 1631–1640.

Hanslmayr, S., & Staudigl, T. (2014). How brain oscillations form memories—A processing based perspective on oscillatory subsequent memory effects. *NeuroImage, 85,* 648–655.

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

Hashimoto, T., Usui, N., Taira, M., & Kojima, S. (2011). Neural enhancement and attenuation induced by repetitive recall. *Neurobiology of Learning and Memory, 96,* 143–149.

Hoechstetter, K., Bornfleth, H., Weckesser, D., Ille, N., Berg, P., & Scherg, M. (2004). BESA source coherence: A new method to study cortical oscillatory coupling. *Brain Topography, 16,* 233–238.

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.

Ille, N., Berg, P., & Scherg, M. (2002). Artifact correction of the ongoing EEG using spatial filters based on artifact and brain signal topographies. *Journal of Clinical Neurophysiology, 19,* 113–124.

Karpicke, J. D., Lehman, M., & Aue, R. W. (2014). Retrieval-based learning: An episodic context account. In B. H. Ross (Ed.), *The psychology of learning and motivation* (Vol. 61, pp. 237–284). San Diego: Elsevier Academic Press.

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

Keresztes, A., Kaiser, D., Kovács, G., & Racsmány, M. (2014). Testing promotes long-term learning via stabilizing activation patterns in a large network of brain areas. *Cerebral Cortex, 24,* 3025–3035.

Kliegl, O., & Bäuml, K.-H.T. (2016). Retrieval practice can insulate items against intralist interference: Evidence from the list-length effect, output interference, and retrieval-induced forgetting. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 42*, 202–214.

Klimesch, W., Doppelmayr, M., Yonelinas, A., Kroll, N. E., Lazzara, M., Röhm, D., & Gruber W. (2001). Theta synchronization during episodic retrieval: neural correlates of conscious awareness. *Cognitive Brain Research, 12*, 33–38.

Klimesch, W., Doppelmayr, M., Pachinger, T., & Russegger, H. (1997). Event-related desynchronization in the alpha band and the processing of semantic information. *Cognitive Brain Research, 6,* 83–94.

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.

Lega, B. C., Jacobs, J., & Kahana, M. (2012). Human hippocampal theta oscillations and the formation of episodic memories. *Hippocampus, 22,* 748–761.

Lehman, M., Smith, M. A., & Karpicke, J. D. (2014). Toward an episodic context account of retrieval-based learning: Dissociating retrieval practice and elaboration. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 40,* 1787–1794.

Maris, E., & Oostenveld, R. (2007). Nonparametric statistical testing of EEG- and MEG-data. *Journal of Neuroscience Methods, 164,* 177–190.

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

Nelson, S. M., Arnold, K. M., Gilmore, A. W., & McDermott, K. B. (2013). Neural signatures of test-potentiated learning in parietal cortex. *The Journal of Neuroscience, 33,* 11754–11762.

Nelson, D. L., McEvoy, C. L., & Schreiber, T. A. (2004). The University of South Florida word association, rhyme, and word fragment norms. *Behavior Research Methods, Instruments, & Computers, 36,* 402–407.

Nyhus, E., & Curran, T. (2010). Functional role of gamma and theta oscillations in episodic memory. *Neuroscience and Biobehavioral Reviews, 34,* 1023–1035.

Otten, L. J., & Rugg, M. D. (2001). Task-dependency of the neural correlates of episodic encoding as measured by fMRI. *Cerebral Cortex, 11,* 1150–1160.

Pashler, H., Cepeda, N. J., Wixted, J. T., & Rohrer, D. (2005). When does feedback facilitate learning of words? *Journal of Experimental Psychology: Learning, Memory, and Cognition, 31,* 3–8.

Pastötter, B., & Bäuml, K.-H. T. (2014). Distinct slow and fast cortical theta dynamics in episodic memory retrieval. *NeuroImage, 94,* 155–161.

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

Pastötter, B., Berchtold, F., & Bäuml, K.-H. T. (2012). Oscillatory correlates of controlled speed-accuracy tradeoff in a response-conflict task. *Human Brain Mapping, 33,* 1834–1849.

Pastötter, B., Dreisbach, G., & Bäuml, K.-H. T. (2013). Dynamic adjustments of cognitive control: Oscillatory correlates of the conflict adaptation effect. *Journal of Cognitive Neuroscience, 25,* 2167–2178.

Pfurtscheller, G., & Aranibar, A. (1977). Event-related cortical desynchronization detected by power measurements of scalp EEG. *Electroencephalography and Clinical Neurophysiology, 42,* 817–826.

Pfurtscheller, G., & Lopes da Silva, F. H. (1999). Event-related EEG/MEG synchronization and desynchronization: Basic principles. *Clinical Neurophysiology, 110,* 1842–1857.

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.

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

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.

Roediger, H. L., & Karpicke, J. D. (2006). Test-enhanced learning: Taking memory tests improves longterm retention. *Psychological Science, 17*, 249–255.

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.

Rosburg, T., Johansson, M., Weigl, M., & Mecklinger, A. (2015). How does testing affect retrieval-related processes? An event-related potential (ERP) study on the short-term effects of repeated retrieval. *Cognitive, Affective, & Behavioral Neuroscience, 15,* 195–210.

Rutishauser, U., Ross, I. B., Mamelak, A. N., & Schuman, E. M. (2010). Human memory strength is predicted by theta-frequency phase-locking of single neurons. *Nature, 464,* 903–906.

Soderstrom, N. C., & Bjork, R. A. (2015). Learning versus performance: An integrative review. *Perspectives on Psychological Science, 10,* 176–199.

Stancák, A., & Pfurtscheller, G. (1996). Event-related desynchronisation of central beta-rhythms during brisk and slow self-paced finger movements of dominant and nondominant hand. *Cognitive Brain Research, 4,* 171–183.

Staudigl, T., & Hanslmayr, S. (2013). Theta oscillations at encoding mediate the context-dependent nature of human episodic memory. *Current Biology, 23,* 1101–1106.

Staudigl, T., Hanslmayr, S., & Bäuml, K.-H. T. (2010). Theta oscillations reflect the dynamics of interference in episodic memory retrieval. *The Journal of Neuroscience, 30,* 11356–11362.

Storm, B. C., Friedman, M. C., Murayama, K., & Bjork, R. A. (2014). On the transfer of prior tests or study events to subsequent study. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 40,* 115–124.

Summerfield, C., & Mangels, J. A. (2005). Coherent theta-band EEG activity predicts item-context binding during encoding. *NeuroImage, 24,* 692–703.

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

Tzagarakis, C., Ince, N. F., Leuthold, A. C., & Pellizzer, G. (2010). Beta-band activity during motor planning reflects response uncertainty. *The Journal of Neuroscience, 30,* 11270–11277.

van den Broek, G. S. E., Takashima, A., Segers, E., Fernández, G., & Verhoeven, L. (2013). Neural correlates of testing effects in vocabulary learning. *NeuroImage, 78,* 94–102.

van Driel, J., Swart, J. C., Egner, T., Ridderinkhof, K. R., & Cohen, M. X. (2015). (No) time for control: Frontal theta dynamics reveal the cost of temporally guided conflict anticipation. *Cognitive, Affective, & Behavioral Neuroscience, 15,* 787–807.

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.

Wing, E. A., Marsh, E. J., & Cabeza, R. (2013). Neural correlates of retrieval-based memory enhancement: An fMRI study of the testing effect. *Neuropsychologia, 51,* 2360–2370.

Wirebring, L. K., Wiklund-Hörnqvist, C., Eriksson, J., Andersson, M., Jonsson, B., & Nyberg, L. (2015). Lesser neural pattern similarity across repeated tests is associated with better long-term memory retention. *The Journal of Neuroscience, 35,* 9595–9602.