**Functional underpinnings of feedback-enhanced test-potentiated encoding**

Petra Ludowicy[1*], Daniela Czernochowski[1], Jaione Arnaez-Telleria[2], Kshipra Gurunandan[2], Thomas Lachmann[1,3], & Pedro M. Paz-Alonso[2,4*]

[1]Center for Cognitive Science, University of Kaiserslautern, Kaiserslautern, Germany;

[2]BCBL. Basque Center on Cognition, Brain and Language, San Sebastian, Spain;

[3]Facultad de Lenguas y Educación, Centro de Investigación Nebrija en Cognición (CINC), Universidad Nebrija, Madrid (Spain);

[4]Ikerbasque, Basque Foundation for Science, Bilbao, Spain.

[*]Correspondence should be addressed to Petra Ludowicy, Center for Cognitive Science, University of Kaiserslautern, 67663, Kaiserslautern, Germany; or to Pedro M. Paz-Alonso, BCBL, Paseo Mikeletegi 69, 2, 20009 Donostia-San Sebastián, Spain. E-mails: pludowicy@gmail.com, kepa.pazalonso@gmail.com

**Highlights**

- Explicit performance feedback boosted memory performance beyond standard test-potentiated encoding
- Repeated study engaged frontoparietal networks, whereas testing recruited lateral temporal regions
- Testing with additional performance feedback strengthened posterior hippocampal engagement relative to studying; testing without feedback was intermediate
- Testing with additional performance feedback boosted functional coupling between hippocampus and regions involved in retrieval and in reward or feedback-related processing

**Abstract**

The testing effect describes the finding that retrieval practice enhances memory performance compared to restudy practice. Prior evidence demonstrates that this effect can be boosted by providing feedback after retrieval attempts (i.e., test-potentiated encoding, TPE). The present fMRI study investigated the neural processes during successful memory retrieval underlying this beneficial effect of correct answer feedback compared with restudy, and whether additional performance feedback leads to further benefits. Twenty-seven participants learned cue-target pairs by I) restudying, II) standard TPE including a restudy opportunity or III) TPE including a restudy opportunity immediately after positive or negative performance feedback. One day later, a cued retrieval recognition test was performed inside the MRI scanner. Behavioral results confirmed the testing effect, and that adding explicit performance feedback enhanced memory relative to restudy and standard TPE. Stronger functional engagement while retrieving items previously restudied was found in lateral prefrontal cortex (PFC) and superior parietal lobe (SPL). In contrast, lateral temporo-parietal areas were more strongly recruited while retrieving items previously tested. Performance feedback increased hippocampal activation and resulted in stronger functional coupling between hippocampus, supramarginal gyrus (SMG) and ventral striatum with lateral temporo-parietal cortex. Our results unveil the main functional dynamics and connectivity nodes underlying memory benefits from additional performance feedback.

Keywords: Learning, Feedback, Testing effect, functional MRI, Hippocampus

## Introduction

Practice makes perfect. Everybody agrees, but there are different ways to enhance long-term memory. A large body of evidence suggests that simple repetition, i.e. restudying, is often preferred by learners, but is less effective than integrating retrieval practice into learning phases (see Roediger & Butler, 2011; Roediger & Karpicke, 2006a,b). For instance, when learning vocabulary in a foreign language, using cue words as prompts to recall the translation from memory (testing) is a more effective learning strategy compared to simply restudying the word together with its translation. This highly reproducible and robust phenomenon is called the "testing effect". In addition, prior research has shown that providing another study opportunity after testing can further increase long-term retention, which has been termed "test-potentiated encoding" (TPE; see Arnold & McDermott, 2013; van den Broek et al., 2016). When learning vocabulary, testing each item, then presenting the word with its translation, can boost subsequent retrieval of the correct word, presumably due to an enhancement of encoding processes.

Several theories and hypotheses have been put forward regarding the mechanisms underlying these effects. Most researchers have focused either on the idea that testing helps to elaborate the semantic network or that testing reinforces the representation of information and thus leads to an improvement of the selection processes during memory retrieval (Rowland, 2014; van den Broek et al., 2016). In line with these theoretical accounts, research investigating the neural basis of the testing effect using functional magnetic resonance imaging (fMRI) reported increased activation in prefrontal cortex (PFC) areas, such as the middle frontal gyrus (MFG) and inferior frontal gyrus (IFG), when practicing by testing compared to restudying in the repetition phase (e.g., Rosner, Elman, & Shimamura, 2014; Vannest et al., 2012). The involvement of these PFC regions has been associated with either higher attentional demands or improved conflict monitoring induced by testing (Rosner et al., 2014; van den Broek, Takashima, Segers, Fernández, & Verhoeven, 2013). In contrast, practicing by restudying compared to testing has been associated with the engagement of temporo-parietal regions, such as the inferior parietal lobe (IPL) and middle temporal gyrus

(MTG), which has been related to accessing semantic representations (van den Broek et al., 2013; Vannest et al., 2012, Wing, Marsh, & Cabeza, 2013).

While these studies have assessed neurocognitive processing in the learning phase, the mechanisms underlying successful retrieval of items encoded via these different routes are poorly understood thus far. When focusing on activation during a final retrieval test (after a delay), previous studies have reported mixed findings ranging from no differences in neural activation for previously tested compared to restudied items (Rosner et al., 2014) to a decrease in neural activation in fronto-parietal networks following practice testing compared to restudying (Keresztes, Kaiser, Kovács, & Racsmány, 2013; Wiklund-Hörnqvist, Stillesjö, Andersson, Jonsson, & Nyberg, 2021). This suggests that practicing by testing reduces the neural activation required to perform accurately at final retrieval (van den Broek et al., 2016). Moreover, recent research on the testing effect focused on the medial temporal lobe (MTL) and reported increased hippocampal engagement due to testing, with anterior hippocampus being more strongly recruited after multiple testing and posterior hippocampus being more engaged for items successfully retrieved only once during practicing (Wiklund-Hörnqvist et al., 2021).

At present, only a few neuroimaging studies have investigated the neural mechanisms underlying learning from correct answer feedback after practice testing (TPE) during the repetition phase, also reporting inconsistent results. In addition to regions listed previously, Liu and colleagues' (2014) study revealed a marginal increase in striatal (caudate, putamen) activation for successful retrieval that was preceded by corrective feedback for incorrect responses (Liu, Liang, Li, & Reder, 2014). Supporting this finding, Wiklund-Hörnqvist, Andersson, Jonsson, and Nyberg (2017) reported increased activation of the ventral striatum during the first correct retrieval, which then decreased in following tests. In contrast, Vestergren and Nyberg (2014) did not find such differences in functional activation patterns when comparing subsequently remembered versus forgotten items.

So far, studies exploring testing with additional correct answer feedback have mostly presented the to-be-learned material in the same way as in the restudy condition (Rowland,

2014). Few studies have explicitly added information regarding the accuracy of the retrieval attempt, either by presenting the words “correct” or “incorrect” (e.g., Jacoby, Wahlheim, & Coane, 2010) or changing the font color to green or red (e.g., Ernst & Steinhauser, 2012). In one such behavioral study of word learning, Ludowicy, Paz-Alonso, Lachmann, & Czernochowski (under review) provided positive or negative feedback indicating test performance on the previously presented item immediately before the correct word was displayed as a restudy opportunity. Results revealed a modest but consistent beneficial effect of the additional performance feedback on retrieval performance in a final test conducted one day after the initial learning. In line with prior studies (e.g., Vestergren & Nyberg, 2014), the authors suggested that their findings might be explained by the relocation of attentional resources and reinforcing aspects of performance feedback as well as elaboration of the semantic network. On correct trials, feedback may diminish internal processes related to self-performance evaluation prior to presentation of the correct answer, which may increase the resources available for semantic elaboration. In contrast, on incorrect trials, this additional feedback information might support attention shifting towards the correct answer feedback and this search-set restriction may enhance encoding processes (Ludowicy et al., under review). However, to the best of our knowledge, no fMRI studies to date have examined the functional mechanisms underpinning the effects of positive and negative feedback on subsequent test-potentiated encoding (i.e., feedback-assisted TPE).

Therefore, the present fMRI study constitutes the first study investigating the neural basis of memory retrieval underlying performance feedback enhanced TPE. As described below, participants were asked to learn word pairs with low association strength in an initial study phase, immediately followed by a practice phase in which the word pairs were studied in three different ways. Participants either restudied the word pairs, were prompted to attempt to retrieve the target word before the correct answer was provided as a restudy opportunity, or attempted retrieval followed by performance feedback and then the correct answer feedback. One day later, participants performed a final test on all previously learned word

pairs inside the MRI scanner. The present study was focused on exploring successful memory retrieval at the final test by comparing the neural activations when retrieving items encoded via different routes (i.e. restudying, testing and testing with performance feedback).

Our study had four goals. First, we expected to replicate the behavioral results observed in our previous paper (Ludowicy et al., under review) as well as the fMRI results for tested compared to restudied materials (e.g., Keresztes et al., 2013; Wiklund-Hörnqvist et al., 2021). Given previous evidence on successful episodic memory retrieval (e.g., Spaniol et al., 2009), we expected to find increased involvement of frontal and parietal areas following restudying compared to testing on correct trials (van den Broek et al., 2016; Wiklund-Hörnqvist et al., 2017; but see Wiklund-Hörnqvist et al., 2021). Moreover, we sought to extend prior findings by comparing items tested following additional performance feedback to those tested with only correct answer feedback and to items which were restudied. Second, we examine differences in anterior versus posterior hippocampus based on accounts suggesting differential involvement of hippocampal regions along the Y-axis (e.g., Poppenk, Evensmoen, Moscovitch, & Nadel 2013; Poppenk & Moscovitch, 2011) with the posterior hippocampus being more involved in retrieval of detailed, pattern-separated representations, hence more involved in test than restudy (see Wiklund-Hörnqvist et al., 2021). Third, to specifically focus on the predicted beneficial effect of performance feedback, we also examine the involvement of ventral striatum due to its well-known recruitment in feedback-related learning (e.g., O'Doherty et al., 2004; Wiklund-Hörnqvist et al., 2017). Fourth, we examine functional coupling between left hippocampus, ventral striatum and other frontoparietal areas reported to be involved in the testing effect and extend it to additional performance feedback (see Wing et al., 2013).

## Methods

*Participants*

A total of 29 native Spanish speakers participated in this study. Data from two participants was excluded due to poor memory performance (2 SD lower than the mean) on the final test after a 1-day retention interval to allow consolidation of learned materials. The

final sample consisted of 27 right-handed participants (mean age = 24.2 years, SD = 4.4 years; 15 females), who reported normal or corrected-to-normal vision, no red-green deficiency and no history of major medical, neurological or psychiatric disorders. The study protocol was approved by the Ethics Committee of the Basque Center on Cognition, Brain and Language (BCBL) and was carried out in accordance with the Code of Ethics of the World Medical Association (Declaration of Helsinki) for experiments involving human participants. Prior to their inclusion in the study, all participants provided informed written consent. Participants received monetary compensation for their participation.

*Materials*

The stimulus material consisted of 180 weakly associated cue-target word pairs (e.g., *feather-duck*; *towel-soap*) taken from the database by Nelson, McEvoy, and Schreiber (2004) and translated into Spanish. All word pairs were chosen based on their forward (FSG) (>.04), backward (BSG) (>.04), mediated (MSG) (>.04) and overlapping (OSG) (>.05) association probability to ensure that participants would need to learn the associations to provide correct responses and were unlikely to produce them by guessing. The word pairs were randomly assigned to one of the repetition practice conditions for each participant.

*Procedure*

The experiment was divided into two sessions spaced approximately 24 hours apart. Figure 1 provides an overview of the experimental design. On Day 1, participants started by intentionally studying all word pairs, followed by three repetition practice cycles in which items were either restudied, tested with a restudy opportunity (TPE) provided by presentation of the correct response, or tested with performance feedback followed by a restudy opportunity (TPE+FB).

On Day 2, participants performed a final test during MRI scanning, followed by a second final test in a behavioral cabin in which they were asked to verbalize all word pairs. This design was chosen to ensure participants could retrieve at least 50% correct word pairs across all repetition practice conditions.

During all cycles, short breaks of 30 s were added after blocks of 45 word pairs and a fixation cross of 1 s preceded every stimulus presentation. Stimuli were presented using Presentation software (Neurobehavioral Systems, Berkeley, CA).

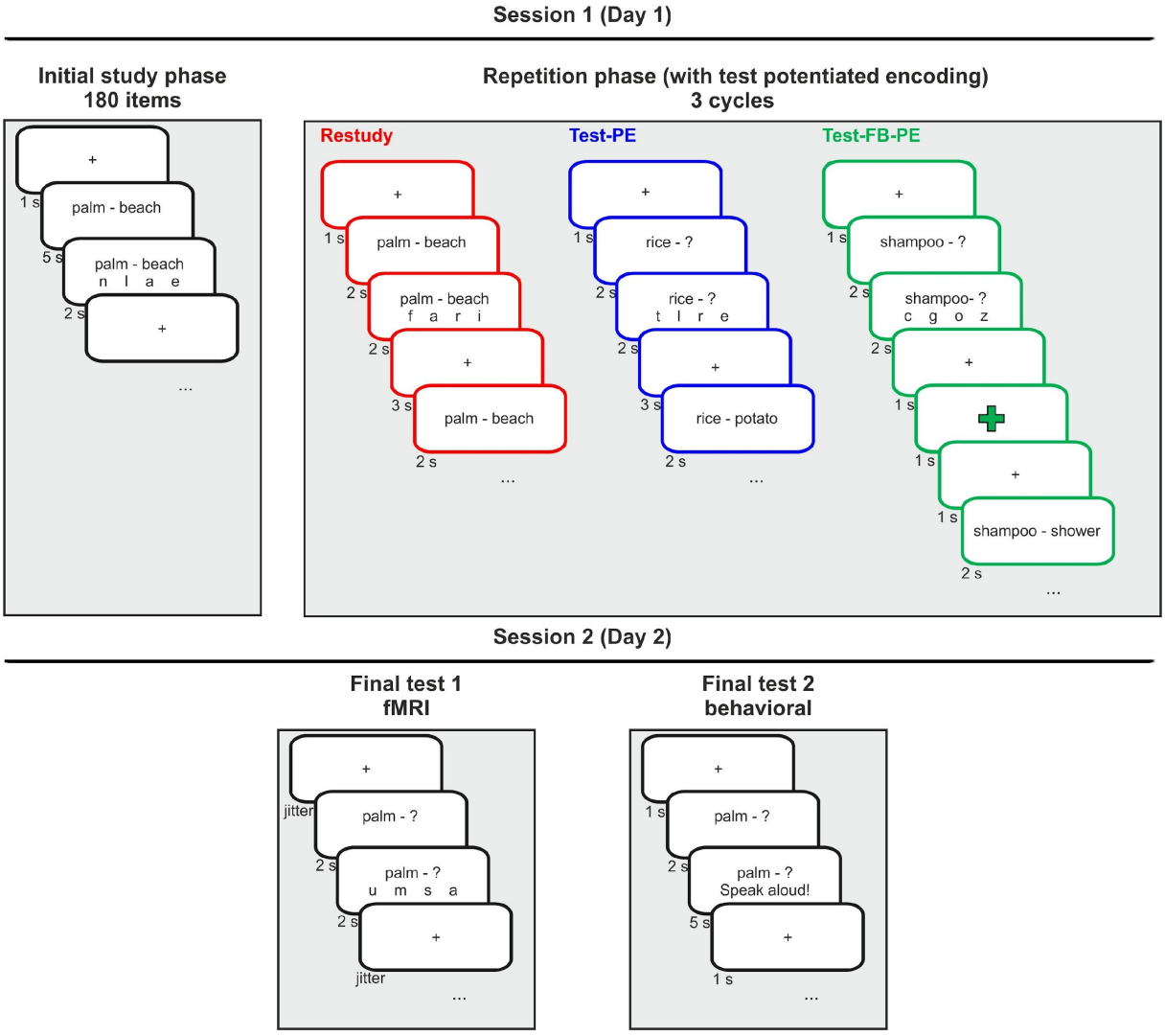



**Figure 1.** Depiction of the experimental procedure.

*Session 1: initial study phase (behavioral)*

In the first phase, participants studied all 180 items, one pair at a time. Each word pair was displayed in random order for a total of 7 s and participants were instructed to try to learn the word pairs for a later test. Participants were informed that the word on the left side was the cue word (e.g., "*palm*") whereas the word on the right was the target word (e.g., "*beach*"), which they would need to retrieve in a later test when presented with the cue word (e.g., "*palm*"). After 5 s, four letters appeared below the stimulus and participants were

instructed to report the third letter of the target word by pressing the corresponding key (e.g., “a” is the third letter of the word “beach”). Participants were instructed to respond within the final 2 s in which the cue word was still present on the screen. This task was added to ensure that participants paid attention to the word pairs and to familiarize them with the response procedure for the next phases, as this task would be used to determine correct responses in the MRI test phase.

*Session 1: repetition practice phase (behavioral)*

In the second part of session 1, participants practiced all 180 word pairs in three repetition cycles, to boost later memory performance. One-third of the word pairs (i.e., 60 items) were randomly assigned to each of the three repetition practice conditions (Restudy, Test-PE, or Test-FB-PE) for each participant. Within each repetition cycle, word pairs were presented randomly such that practice conditions were intermingled.

In the *restudy condition*, the cue and target were presented together for 4 s. In the two test conditions, only the cue word was presented for 4 s, with a “?” instead of the target word. As in the initial study phase, on all trials four letters appeared on the screen during the last 2 s of word presentation and participants were requested to indicate the third letter of the target word. The correct response letter was presented randomly at one of the four possible positions below the study or test item and the distractors were randomly selected from a set of possible response letters. To ensure that participants were unable to use the response options as cues for memory retrieval, they were changed for each study or test cycle. In the test conditions, participants were instructed to provide their best guess if they did not know the target word. A covert response procedure via button press was selected to minimize motion artifacts and to allow for the assessment of reaction times (see also e.g., Pastötter & Bäuml, 2016; van den Broek et al., 2013, 2014; Wiklund-Hörnqvist et al., 2021).

In the testing conditions, following the 4 s presentation of the cue word, the cue and target word were presented together for 2 s to provide an opportunity to re-encode the items (i.e., test-potentiated encoding, TPE). Prior to this correct answer feedback, in the *Test-PE condition*, a fixation cross was displayed on the screen for 3 s. In the *Test-FB-PE condition*,

performance feedback was provided for 1 s, preceded and followed by a fixation cross for 1 s each. A green plus sign was presented if the correct response was given within the 2 s time limit. Conversely, a red minus sign was displayed to indicate incorrect responses.

*Session 2: final test I (fMRI) and test II (behavioral)*

Approximately 24 hours after Session 1, participants returned to the lab for the final test, which was split into two parts. The first test was performed inside the MR scanner, whereas the second one was performed behaviorally outside the MR scanner.

In final test I, the cue words were presented for 4 s with a "?" on the right side. After 2 s, 4 letter options were displayed on the screen and participants were asked to indicate the third letter of the target word via button press. No feedback was provided afterwards. Between consecutive trials, a fixation cross was shown, with a jittered duration. Optimal sequencing software (OptSeq2, https://surfer.nmr.mgh.harvard.edu/optseq/) was employed to determine the variable duration of the jitter fixation (0-5000ms) and the order of trial types in each of the repetition practice conditions within each functional run to optimize efficient recovery of the blood-oxygen level dependent (BOLD) response (Fischl, Sereno, & Dale, 1999). This test was split into two runs with 90 word pairs in each.

Participants were not asked to provide the target words in this covert retrieval test to reduce articulatory artifacts in the functional data during retrieval as a result of overt verbalization. Therefore, to confirm correct responses provided inside the MR scanner, an overt recall task, final test II, was performed outside the MR scanner. As before, the cue word was presented next to a "?" for 2 s. Afterwards, participants were prompted to verbalize the full word pair while the cue word remained on the screen for a further 5 s. Participants were informed that they should respond before the cue word disappeared. Responses were recorded and were rated as correct if the participants correctly recalled the target word within the 5 s time limit. A fixation cross was displayed for 1 s between trials.

*fMRI data acquisition*

Whole-brain MRI data acquisition was conducted on a 3-T Siemens Prisma Fit whole-body MRI scanner (Siemens Medical Solutions) using a 64-channel whole-head coil. Foam

pillows were provided to help stabilize head position and scanner noise was reduced with earplugs and padded headphones. The MRI acquisition included T1-weighted structural images and functional T2* images. High-resolution MPRAGE T1-weighted structural images were collected with the following parameters: time-to-repetition (TR) = 2530 ms, time-to-echo (TE) = 2.36 ms, flip angle (FA) = 7°, field of view (FoV) = 256 mm, voxel resolution = 1 $mm^3$, 176 slices. Functional images were acquired in two consecutive functional runs using a single gradient-echo echo-planar multiband pulse sequence with the following acquisition parameters: TR = 1000 ms; TE = 35 ms; MB acceleration factor = 5; 65 axial slices with a 2.4 $mm^3$ voxel resolution; no inter-slice gap; FA = 56º; FoV = 210 mm; 486 volumes. The first twelve functional volumes in each run were discarded to allow for T1-equilibration effects.

*fMRI data preprocessing*

Standard SPM12 (The Wellcome Department of Cognitive Neurology, London, UK) preprocessing routines and analysis methods were employed. Images were first corrected for differences in timing of slice acquisition and then realigned to the first and mean volumes using rigid-body registration. Each subject's functional volumes were spatially smoothed with a 4-mm full-width half-maximum (FWHM) Gaussian kernel. Motion parameters extracted from the realignment process were used to identify outlier volumes with sudden scan-to-scan motion exceeding 0.5 mm and volumes whose global intensity fluctuated more than 1.3%. These were corrected via interpolation between the nearest non-repaired scans (ArtRepair version 5b; Mazaika, Hoeft, Glover, & Reiss, 2009). After volume repair, functional volumes were co-registered to the T1 images using 12-parameter affine transformation and spatially normalized to the MNI space by applying non-linear transforms estimated by deforming the MNI template to each individual's structural volume. During normalization, the volumes were sampled to 3-mm cubic voxels. The resulting volumes were then spatially smoothed with a 7-mm FWHM Gaussian kernel. Due to the quadratic relation between separate smoothing operations, the total smoothing applied to the functional data was approximately equivalent to smoothing with an 8-mm FWHM Gaussian kernel. Finally, time series were temporally

filtered to eliminate contamination from slow frequency drift (high-pass filter with cut-off period of 128 s).

*fMRI data analyses*

Statistical analyses were performed on individual participant's data using the general linear model (GLM). The motion parameters for translation (x, y, z) and rotation (yaw, pitch, roll) were also included as covariates of noninterest in this GLM. fMRI time series data were modeled by a series of events convolved with a canonical hemodynamic response function (HRF). Two different GLM models were constructed. The first model (i.e., Analysis I), included 3 regressors of interest related to their prior repetition practice condition during the repetition phase: *Restudy*, *Test-PE* and *Test-FB-PE*. These regressors were time-locked to the onset of the stimulus presentation during the final in-scanner memory test. Trials incorrectly retrieved at the final test I were modeled separately and excluded from the main analysis. For the second model (i.e., Analysis II) we conditioned final test results time-locked to the onset of the stimulus presentation during the final in-scanner memory test for items previously tested with and without performance feedback during the repetition practice phase dependent on retrieval performance in repetition cycle 1. Hence, Analysis II included 2 regressors of interest for the conditions Test-PE and Test-FB-PE and 2 regressors of interest for correct and incorrect memory retrieval in repetition cycle 1, respectively. In Analysis II, trials incorrectly retrieved in the final in-scanner memory test as well as trials belonging to the Restudy condition (i.e., previously studied during session 1) were modeled separately and excluded from further analyses.

SPM12 FAST was used for temporal autocorrelation modeling in these GLMs due to its optimal performance in terms of removing residual autocorrelated noise in first-level analyses (Olszowy, Aston, Rua, & Williams, 2019). The least-squares parameter estimates of the height of the best-fitting canonical HRF for each condition were used in pairwise contrasts. Contrast images computed on a participant-by-participant basis were submitted to group analysis. At the group level, whole-brain contrasts between repetition practice conditions were computed by performing one-sample t-tests on these images, treating

participants as a random effect. For Analysis I, a general All > Null (fixation baseline) contrast was computed to identify all the brain voxels involved in this fMRI experimental design across the three repetition practice conditions for all the behaviorally-confirmed in-scanner correct responses. In addition, also for Analysis I, we performed contrasts across all participants for the main conditions of interest: Restudy vs. Test-PE, Restudy vs. Test-FB-PE, Test-PE vs. Test-FB-PE. Familywise error rate (FWE) correction was applied to whole-brain maps involving all participants, with the cluster level set at $p < 0.05$ and using a voxel-extent threshold of $p < 0.001$. Brain coordinates throughout the text, as well as in tables and figures, are reported in MNI atlas space (Cocosco, Kollokian, Kwan, & Evans, 1997).

The MARSBAR toolbox (Brett, Anton, Valabregue, & Poline, 2002) was used for region of interest (ROI) analyses. Given that previous meta-analyses on successful episodic memory retrieval (Spaniol et al., 2009; Neurosynth.org) do not include all the relevant regions that previous studies on the testing effect have reported (van den Broek et al., 2016), we functionally defined left-lateralized ROIs based on the overlap of the general activation produced by our design (All vs. Null) and anatomical masks of regions found to be relevant in previous testing effect studies, including MFG (center of mass= -38 26 34; volume = 13488 mm; e.g., Vannest et al., 2012; Wing et al., 2013), IFG (i.e., *pars opercularis and pars triangularis*; center of mass= -45 21 16; volume = 21288 mm; e.g., van den Broek et al., 2013), SPL (center of mass= -24 -61 57; volume = 15128 mm; e.g., Liu & Reder, 2016; Rosner et al., 2014), SMG (center of mass= -55 -26 32; volume = 3080 mm; e.g., Nelson, Arnold, Gilmore, & McDermott 2013; Vestergren & Nyberg, 2014) and MTG (center of mass= -48 -59 3; volume = 4560 mm; e.g., Wiklund-Hörnqvist et al., 2021).

Since weakly associated word pairs were used as learning material in the present experimental design, we targeted left hemisphere regions in our analyses due to their prominent role in language processing. Moreover, the testing effect has been proposed to predominantly modulate neural activity in the left hemisphere (see Jonsson, Wiklund-Hörnqvis, Stenlund, Andersson, & Nyberg, 2021; Wiklund-Hörnqvist et al., 2021), which also supports this focus. The left ventral striatum (VS) was also included for three reasons: (I)

Previous studies have highlighted its importance for learning from positive and negative feedback (see Costa, Dal Monte, Lucas, Murray, & Averbeck, 2016; O'Doherty et al., 2004) as well as (II) learning from practice testing, especially when combined with correct answer feedback (see Clos, Bunzeck, & Sommer, 2019; Wiklund-Hörnqvist, et al., 2017; center of mass= -25 -4 -6; volume = 1280 mm3). Furthermore, the ventral striatum (III) contributes to declarative memory retrieval and supports memory reactivation of motivationally relevant information (see Lansink et al., 2008; Scimeca, & Badre, 2012). In addition, we included the left anterior and posterior hippocampus, as investigated in prior studies on the testing effect (Wiklund-Hörnqvist, et al., 2021). In contrast to the other ROIs, the hippocampal ROIs were defined structurally due to its specific size and shape.

Functional connectivity analyses were conducted using the beta-series correlation method (Rissman, Gazzaley, & D'Esposito, 2004) implemented in SPM12 with custom MATLAB scripts. The canonical HRF in SPM was fitted to each occurrence of each repetition practice condition and the resulting parameter estimates (beta values) were sorted according to the practice conditions to produce a condition-specific beta series for each voxel. Pairwise functional connectivity analyses were performed for left anterior and posterior hippocampus, left ventral striatum and left SMG with each of the remaining ROIs for each participant and repetition practice condition. Since the correlation coefficient ranges from –1 to +1, an arc-hyperbolic tangent transform (Fisher, 1921) was applied to these beta-series correlation values (r values) to make its null hypothesis sampling distribution approach that of the normal distribution. The normally distributed Fisher's Z values for each ROI pair were submitted to statistical analyses (see below).

*Statistical Analysis*

Day 1 results from the three repetition practice cycles, including retrieval accuracy (% correct) and median reaction times (RT), were analyzed using a 2 x 3 repeated measures ANOVA with the factors *Condition* (Test-PE vs. Test-FB-PE) and *Repetition Cycle* (cycle 1-3).

Day 2 results, including memory performance, RT, functional activation, and connectivity results, were analyzed in two separate analyses: Analysis I and II.

These analyses were focused on successful memory retrieval based on verbalized responses in the final test II. For *Analysis I*, final test results of successfully retrieved items, i.e., memory performance, RT, functional activation and connectivity results, were examined dependent on the repetition practice conditions (i.e. Restudy, Test-PE or Test-FB-PE) and then submitted to a one-way analysis of variance (ANOVA) with post-hoc t-tests. For *Analysis II*, items previously tested with or without feedback were conditioned based on participants' retrieval performance in repetition cycle 1 using a 2 x 2 repeated-measures (rm) ANOVA with the factors *Condition* (Test-PE vs. Test-FB-PE) and *Performance in repetition cycle 1* (correct vs. incorrect retrieval). Performance in repetition cycle 1 was selected for this analysis, since this was the first testing phase in which participants attempted to retrieve the target word and, in the Test-FB-PE condition, received positive or negative performance feedback, which was predicted to influence final test performance. Items belonging to the Restudy condition during the repetition phase could not be compared to items previously tested as participants were always presented with the correct response in the repetition phase (the cue and the target words were presented simultaneously). Additionally, Analyses I and II involving the left hippocampus were extended by the factor *Y-axis* (anterior vs. posterior) based on our hypotheses. Finally, we examined the relationship between the neural and behavioral findings by calculating correlational analyses between functional activation and connectivity results and final test recall performance. All ANOVAs, post-hoc t-tests and correlations reported in the results were corrected for multiple comparisons using false-discovery rates (FDR) with a significance level of $q < 0.05$.

## Results

### Behavioral results

On Day 1, all participants first studied and then practiced the word pairs three times by either restudying, testing followed by a restudy opportunity or testing followed by

performance feedback then restudy. We analyzed retrieval performance during the repetition practice phase for items tested either with or without additional feedback with a 3 (*Repetition Cycle*: cycle 1 – 3) X 2 (*Condition*: Test-PE vs. Test-FB-PE) repeated measures ANOVA. Trials belonging to the restudy condition were not included because there was no retrieval. Results revealed statistically significant main effects of *Repetition Cycle* ($F(2, 52) = 152.91$, $q < 0.001$, $\eta_p^2 = 0.86$), which were consistent across all pairwise comparisons ($q$s < 0.001), showing that retrieval performance improved with each cycle (Figure 2, panel A). The effect of *Condition* was also significant (Test-PE vs. Test-FB-PE; $F(1, 26) = 5.08$, $q = 0.049$, $\eta_p^2 = 0.16$), indicating an improvement in retrieval performance for items tested with additional performance feedback. No interaction effect emerged in this analysis ($F < 1$, $q = 0.843$, $\eta_p^2 = 0.007$). The results of the RT analysis for correct responses were highly similar. A statistically significant main effect of *Repetition Cycle* revealed a consistent decrease in RT from cycles 1 to 3 ($F(2, 52) = 61.44$, $q < 0.001$, $\eta_p^2 = 0.70$; all pairwise comparisons: $q < 0.001$) whereas neither the main effect of *Condition* ($F < 1$, $q = 0.686$, $\eta_p^2 = 0.021$) nor the interaction ($F < 1$; $q = 0.932$, $\eta_p^2 = 0.002$) reached statistical significance.

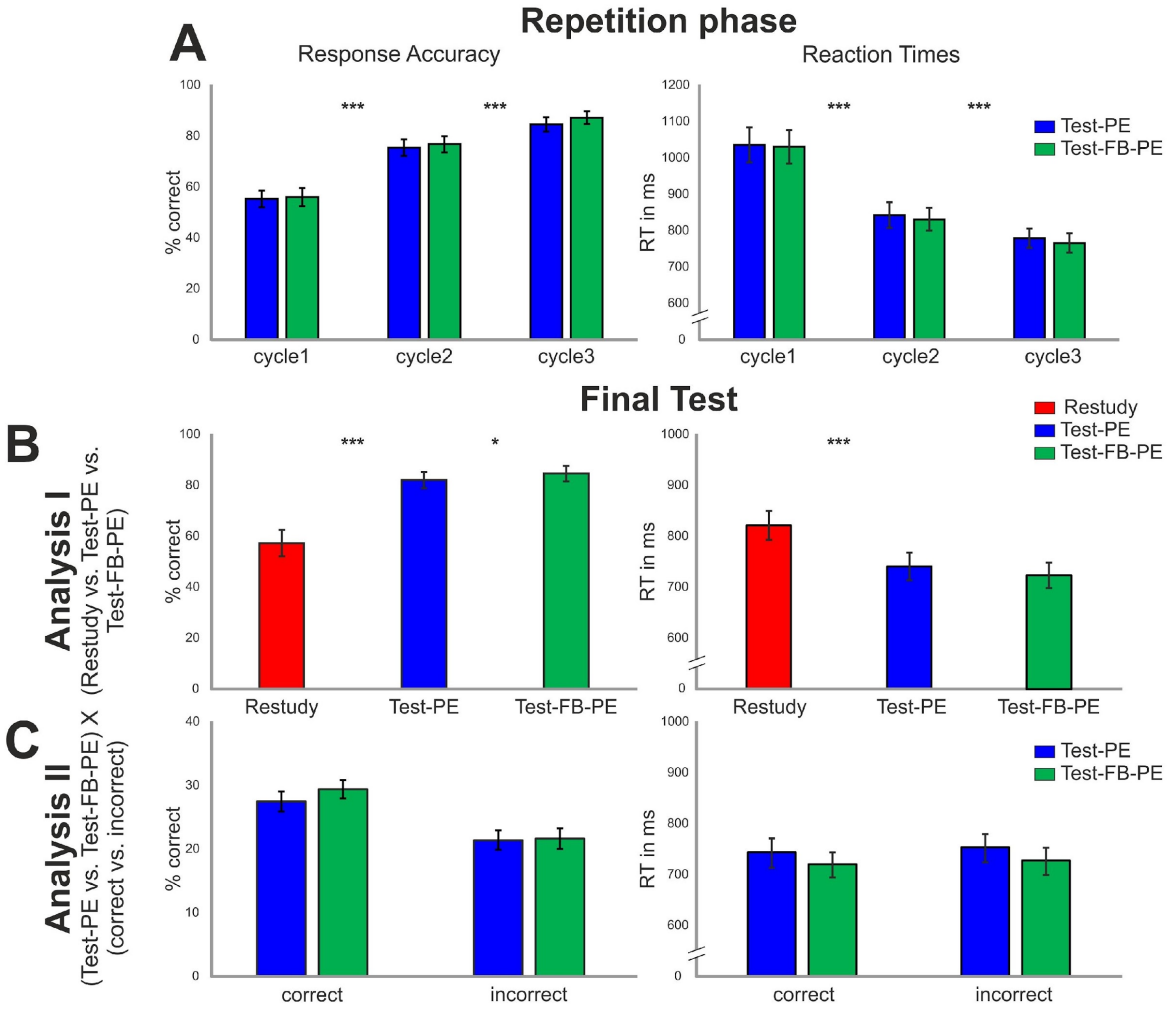



**Figure 2.** Mean percent response accuracy and reaction time results at (A) repetition cycles 1-3 comparing trials tested with or without performance feedback (Test-PE vs. Test-FB-PE). (B) Analysis I at final test comparing the main repetition practice conditions, and (C) Analysis II at final test of items tested with or without performance feedback conditioned on retrieval performance in repetition cycle 1. Error bars show the standard error and asterisks denote statistically significant effects: *$q$ < 0.05, **$q$ < 0.01 ***$q$ < 0.001.

*Analysis I.* On Day 2, all word pairs were tested twice, first in the MRI collecting responses via button presses and afterwards in a behavioral cabin in which participants were asked to say their responses aloud (i.e., the target of the learnt word pairs). Participants retrieved 75% (SE = 4) of all word pairs. The experiment was designed to allow participants to reach high levels of performance to have sufficient correct trials for fMRI analyses. A one-way ANOVA comparing final test retrieval performance for items previously restudied compared to tested (i.e., Test-PE) or tested with performance feedback (i.e., Test-FB-PE; Figure 2, panel B) revealed a statistically significant main effect of *Condition* (i.e., Test-FB-PE; $F(2, 52) = 74.01$, $q < 0.001$, $\eta_p^2 = 0.74$; Restudy vs. Test-PE: $q < 0.001$; Restudy vs.

Test-FB-PE: $q < 0.001$; Test-PE vs. Test-FB-PE: $q = 0.012$). Participants retrieved 57 % of word pairs previously restudied during the repetition practice phase, whereas they reached 82 % for items previously tested and 85 % when additional performance feedback was provided. In line with prior findings regarding the testing effect on RTs, a statistically significant main effect of *Condition* revealed faster responses in the final test for items previously tested (i.e., Test-PE) compared to items restudied in the repetition practice phase, whereas additional performance feedback did not significantly reduce final test RTs beyond the testing effect (i.e., Test-FB-PE; $F(2, 52) = 16.08$, $q < 0.001$, $\eta_p^2 = 0.38$), Restudy vs. Test-PE: $q = 0.001$; Restudy vs. Test-FB-PE: $q < 0.001$; Test-PE vs. Test-FB-PE: $q = 0.091$). Differences between Test-PE and Test-FB-PE might have been masked due to the specific design of the present study, since a covert response procedure was selected to minimize motion artifacts and the response options were presented 2 s after the retrieval test. Hence, participants were able to retrieve the target word and decide the response letter before the response options appeared.

*Analysis II.* To investigate if retrieval performance on Day 1 influenced final test results, we conditioned final test results of items tested either with or without additional performance feedback during the repetition practice phase dependent on retrieval performance in repetition cycle 1. The rm-ANOVA with the Factors *Condition* (Test-PE vs. Test-FB-PE) and *Performance in repetition cycle 1* (correct vs. incorrect) did not reveal any significant differences in response accuracy or RT (response accuracy: *Condition* ($F = 4.49$, $q = 0.066$, $\eta_p^2 = 0.015$); *Performance in repetition cycle 1* ($F = 5.57$, $q = 0.066$, $\eta_p^2 = 0.018$); *Condition* X *Performance in repetition cycle 1* ($F = 1.10$, $q = 0.304$, $\eta_p^2 = 0.041$) / RT: *Condition* ($F = 5.68$, $q = 0.074$, $\eta_p^2 = 0.179$); *Performance in repetition cycle 1* ($F < 1$, $q = 0.734$, $\eta_p^2 = 0.734$); *Condition* X *Performance in repetition cycle 1* ($F < 1$, $q = 0.906$, $\eta_p^2 = 0.001$).

## fMRI results

### Whole-brain contrasts

*Analysis I.* First, we sought to identify brain regions relevant for memory retrieval by contrasting All > Null or fixation baseline. Consistent with previous evidence, the results obtained in the current study revealed activation mostly on the left, but also some right-lateralized activations (see Figure 3). Next, fMRI data were examined based on the repetition practice on Day 1 in order to dissociate brain regions specifically activated for items previously either restudied or tested or tested with additional performance feedback (i.e., Restudy vs. Test-PE vs. Test-FB-PE). All analyses were restricted to items correctly retrieved at the final test. A one-way ANOVA investigating the activation pattern for the three repetition practice conditions revealed no significant differences in neural activation due to testing with or without feedback (i.e., Test-PE vs. Test-FB-PE) and therefore, the results of the whole-brain analysis were examined together for Test-PE and Test-FB-PE conditions. Brain areas more activated for previously restudied compared to tested items (i.e., Restudy > Test-PE/Test-FB-PE) included left-lateralized regions: *pars triangularis, pars orbitalis*, and MFG in the PFC, precentral gyrus, inferior temporal gyrus, SPL and inferior and middle occipital gyrus extending to fusiform gyrus. In contrast, correct retrieval of items previously tested either with or without performance feedback (i.e., Restudy < Test-PE/Test-FB-PE) resulted in increased activation in bilateral temporoparietal regions, including superior temporal gyrus (STG), rolandic operculum, precentral gyrus and postcentral gyrus as well as left SMG and right temporal pole (Figure 3, Table S1).

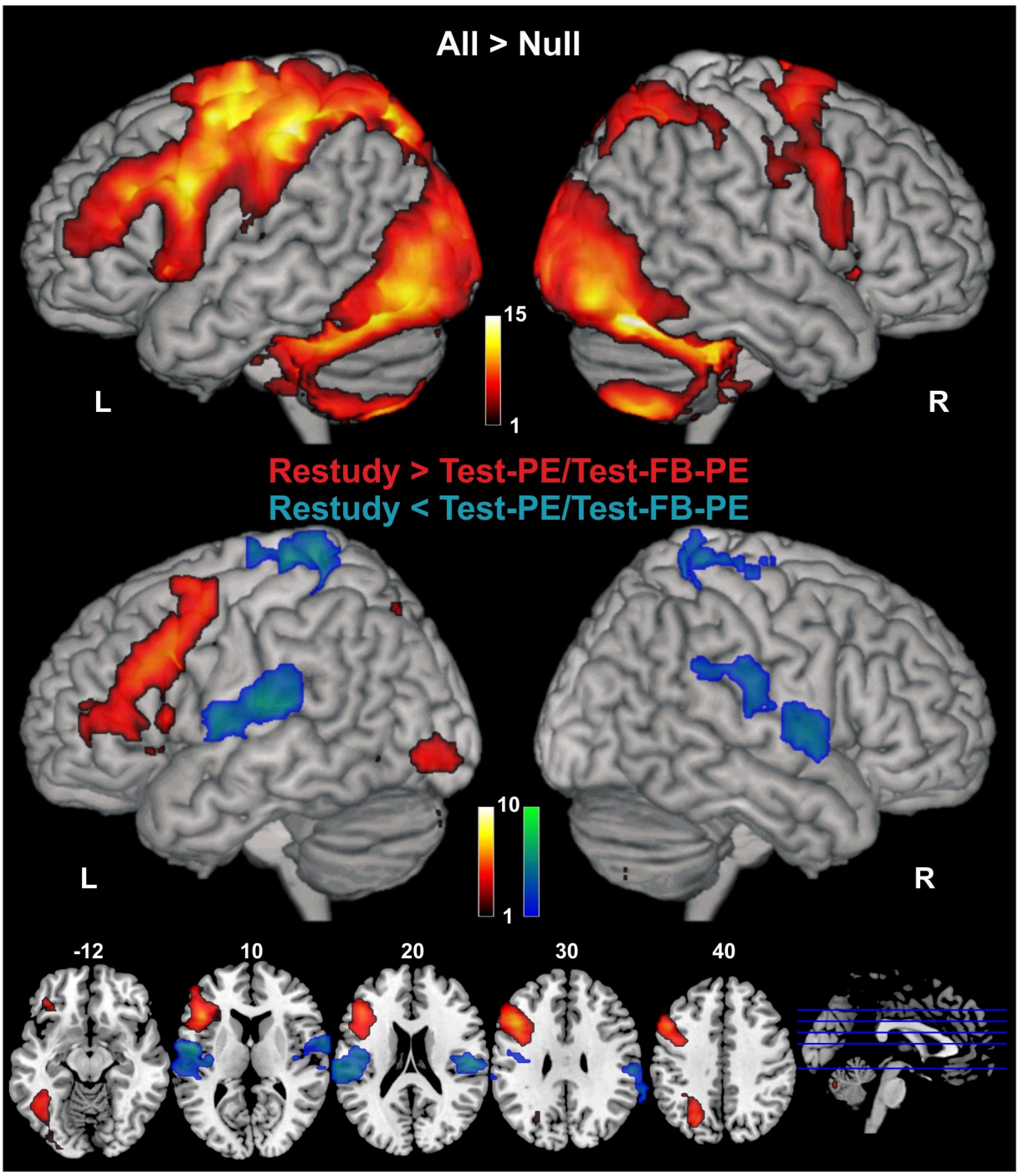



**Figure 3.** Brain renderings and axial slices showing activations for the All > Null whole-brain contrast, as well as for the specific contrasts Restudy > Test-PE/Test-FB-PE (red-yellow colors) and Restudy < Test-PE/Test-FB-PE (blue-green colors), across all subjects using a voxel-level significance threshold set at $p < 0.001$ and a FWE-corrected cluster level significance threshold set at $p < 0.05$.

**ROI analyses**

*Analysis I.* To evaluate regional activation patterns based on the practice conditions (i.e., Restudy, Test-PE, Test-FB-PE), we conducted ROI analyses on regions reported in previous studies and meta-analyses. These regions included left hemisphere MFG, IFG,

SPL, SMG, MTG and ventral striatum (Figure 4 A). The fMRI parameter estimates were extracted for each repetition practice condition of interest against baseline for each of the ROIs. Percent signal change values for each ROI were submitted separately to one-way ANOVAs with *Condition* (Restudy vs. Test-PE vs. Test-FB-PE) as the within-subject factor. Results revealed increased activation for previously restudied compared to tested word pairs in left MFG, left IFG and left SPL (Table 1, Figure 4 A). In contrast, left SMG was more strongly engaged for previously tested items relative to restudied items. No differences were found due to additional performance feedback. Results for the ventral striatum ($F$ < 1, $q$ = 0.519, $\eta_p^2$ = 0.025) and MTG ($F$ = 1.675, $q$ = 0.225, $\eta_p^2$ = 0.061) did not reveal statistically significant differences.

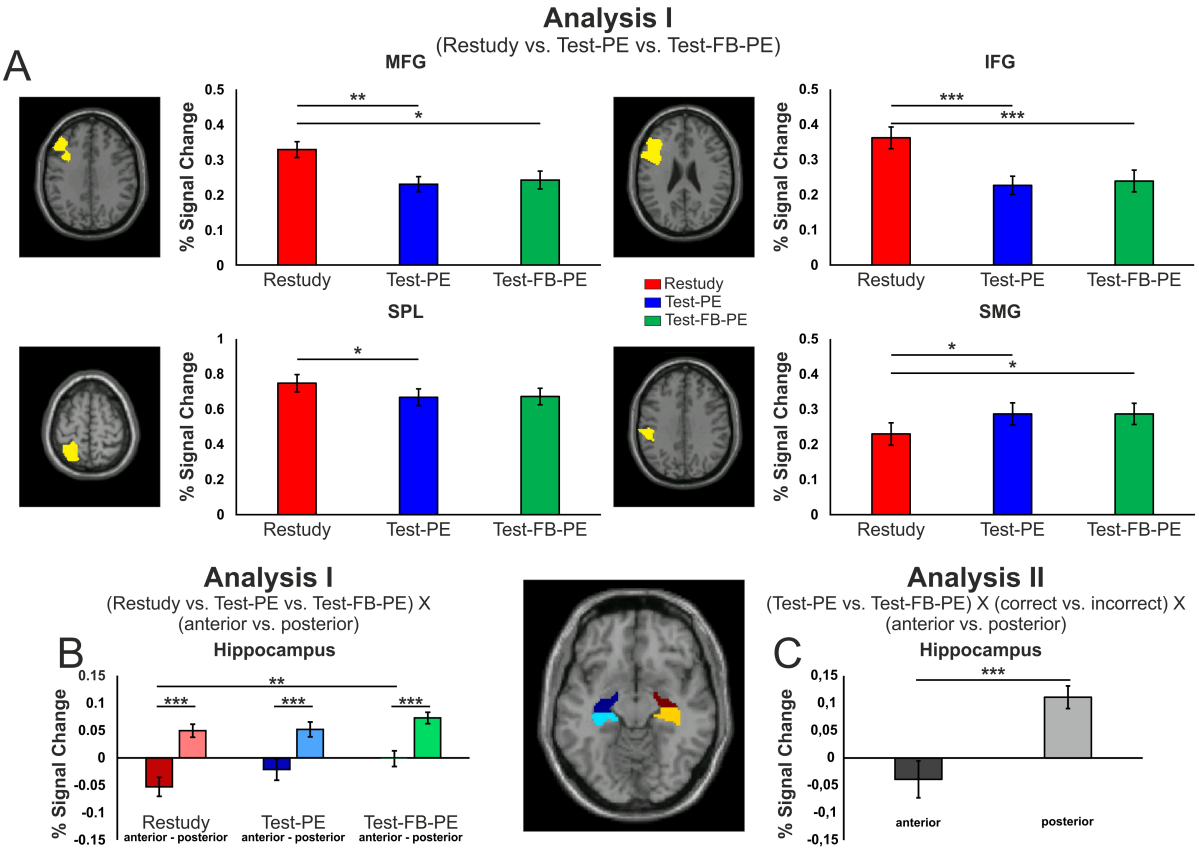



**Figure 4.** ROI analyses for Analysis I (A & B) and II (C). (A) Regions in Analysis I revealing increased % signal change for Restudy > Test-PE/Test-FB-PE in MFG, IFG and SPL; and for Restudy < Test-PE/Test-FB-PE in SMG. (B) Hippocampal activation pattern in Analysis I revealing increased % signal change for Test-FB-PE > Restudy. (C) Hippocampal activation pattern in Analysis II showing increased % signal change in posterior (lighter colors) versus anterior (darker colors) hippocampus based on retrieval performance in repetition cycle 1. Asterisks indicate statistically significant effects: *$q$ < 0.05, **$q$ < 0.01 ***$q$ < 0.001. MFG = middle frontal gyrus; IFG = inferior frontal gyrus; SPL = superior parietal lobe; SMG = supramarginal gyrus.

**Table 1: Summary of ROI Results for Analysis I (Restudy vs. Test-PE vs. Test-FB-PE)**

| ROI | F-values | $\eta_p^2$ | FDR q-values | Post-hoc tests (FDR q-values) | | |
|---|---|---|---|---|---|---|
| | | | | Restudy vs. Test-PE | Restudy vs. Test-FB-PE | Test-PE vs. Test-FB-PE |
| L. MFG | F(2, 50) = 9.25 | 0.27 | **0.003** | **0.003** | **0.010** | 0.501 |
| L. IFG | F(2, 52) = 21.76 | 0.46 | **< 0.001** | **< 0.001** | **< 0.001** | 0.555 |
| L. SPL | F(2, 52) = 4.88 | 0.16 | **0.031** | **0.036** | 0.056 | 0.835 |
| L. SMG | F(2, 52) = 4.73 | 0.15 | **0.024** | **0.028** | **0.027** | 0.995 |

In line with recent literature (Wiklund-Hörnqvist et al. 2021) providing evidence for different activation in the anterior and posterior hippocampus, we also investigated the left anterior and posterior hippocampus. A 3 x 2 repeated measures ANOVA with the factors *Condition* (Restudy vs. Test-PE vs. Test-FB-PE), and *Y-axis* (anterior vs. posterior) revealed main effects of *Condition* ($F(2, 48) = 4.43$, $q = 0.025$, $\eta_p^2 = 0.16$), and *Y-axis* ($F(1, 24) = 39.47$, $q < 0.001$, $\eta_p^2 = 0.62$). The interaction of *Condition* X *Y-axis* was also statistically significant ($F(2, 48) = 6.84$, $q = 0.005$, $\eta_p^2 = 0.22$; Restudy anterior vs. Restudy posterior: $q < 0.001$; Test-PE anterior vs. Test-PE posterior: $q < 0.001$; Test-FB-PE anterior vs. Test-FB-PE posterior: $q < 0.001$; Restudy anterior vs. Test-PE anterior: $q = 0.084$; Restudy anterior vs. Test-FB-PE anterior: $q = 0.002$; Test-PE anterior vs. Test-FB-PE anterior: $q = 0.204$; Restudy posterior vs. Test-PE posterior: $q = 0.854$; Restudy posterior vs. Test-FB-PE posterior: $q = 0.082$; Test-PE posterior vs. Test-FB-PE posterior: $q = 0.084$; see Figure 4 B).

*Analysis II.* Next, ROI results for word pairs tested with or without additional performance feedback (i.e., Test-PE vs. Test-FB-PE) were conditioned on the participant's performance during the first repetition cycle on Day 1. Similar to the behavioral final test results using a 2 x 2 repeated measures ANOVA with the factors *Condition* (Test-PE vs. Test-FB-PE) and *Performance in repetition cycle 1* (correct vs. incorrect), here all previously described ROIs were analyzed, but none revealed significant differences ($Fs \leq 2.53$, $qs \geq 0.511$, $\eta_{ps}^2 \leq 0.089$).

For the left hippocampus, a 2 x 2 x 2 repeated measures ANOVA with the factors *Condition* (Test-PE vs. Test-FB-PE), *Performance in repetition cycle 1* (correct vs. incorrect) and *Y-axis* (anterior vs. posterior) was conducted, revealing only a significant main effect for *Y-axis* ($F(1, 25) = 31.57$, $q < 0.001$, $\eta_p^2 = 0.56$) (see Figure 4 C). The same pattern of results was observed when considering both the left and right hippocampus in *Analysis I* and *II*.

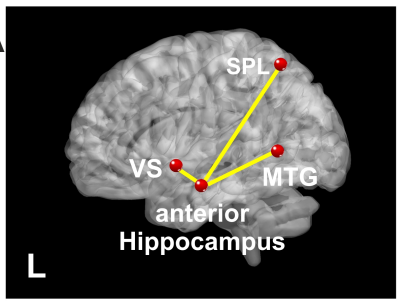
**Functional connectivity analyses**

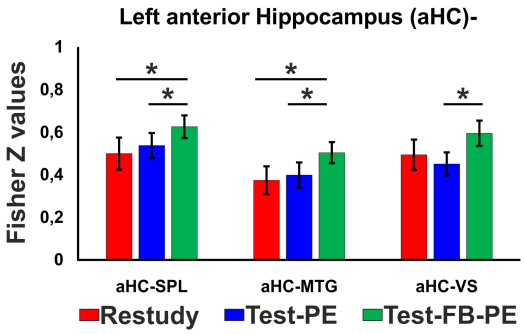
*Analysis I.* Prior fMRI evidence examining the testing effect suggested that testing influences functional coupling between the hippocampus and other brain regions (Wing et al., 2013). Here we investigated functional connectivity patterns of the left anterior and posterior hippocampus with the remaining regions of interest previously examined (i.e., MFG, IFG, MTG, SPL, SMG, ventral striatum). In addition, as the SMG was the only region that showed increased engagement for testing compared to studying in Analysis I, functional connectivity between left SMG with the remaining regions of interest was analyzed. Finally, due to the central role of the ventral striatum for feedback processing and its well-known anatomical connections with the hippocampus (see Kahn & Shohamy, 2013; Lisman & Grace, 2005), functional coupling of the ventral striatum with the remaining regions of interest was also examined. Hence, for Analysis I, a one-way ANOVA with the factor *Condition* (Restudy vs. Test-PE vs. Test-FB-PE) was conducted using condition-specific Fisher's *Z* transformed beta-series values (see Figure 5A). Results revealed a main effect of *Condition* for the functional coupling of the left anterior hippocampus (aHC) with MTG, SPL and ventral striatum. Results are reported in Table 2.

# Analysis I

(Restudy vs. Test-PE vs. Test-FB-PE)

A

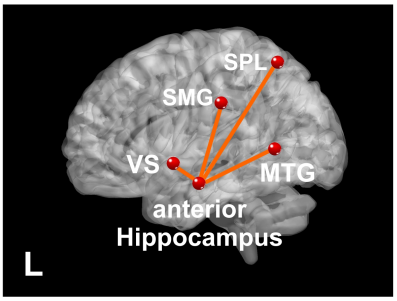



# Analysis II

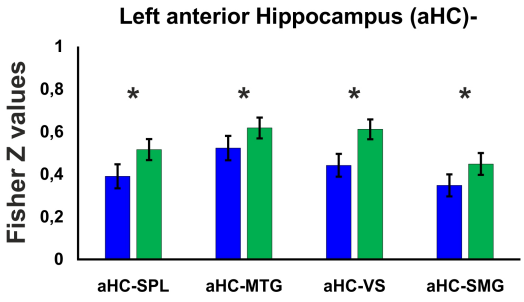
(Test-PE vs. Test-FB-PE) X (correct vs. incorrect)

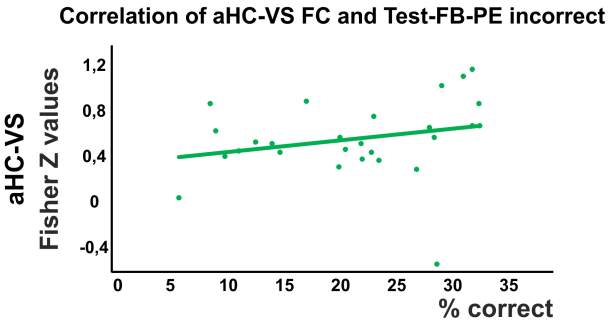
B



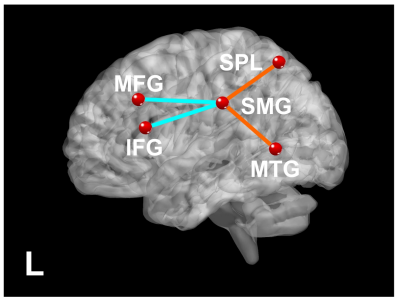

Test-PE Test-FB-PE

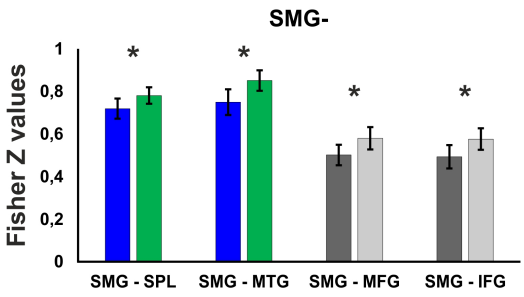
correct incorrect



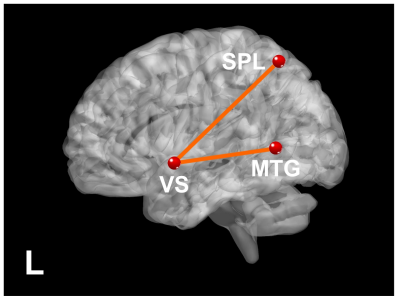




C




D



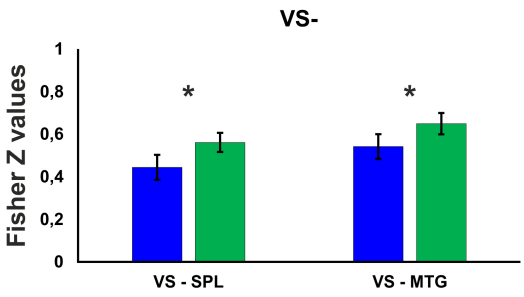



**Figure 5.** Pairwise functional connectivity analyses for Analysis I (A) and Analysis II (B, C & D). (A) Functional coupling of left aHC-SPL, aHC-MTG and aHC-VS was increased for items tested with additional performance feedback. (B) Increased aHC-SPL, aHC-MTG, aHC-VS and aHC-SMG functional coupling for Test-FB-PE versus Test-PE; also, brain-behavior correlations: aHC-VS functional coupling increased with retrieval success only for items previously tested with additional performance feedback and retrieved incorrectly in repetition cycle 1. (C) Increased SMG-SPL, SMG-MTG functional coupling for Test-FB-PE versus Test-PE; also, increased SMG-MFG and SMG-IFG functional coupling for incorrectly versus correctly retrieved items in repetition cycle 1. (D) Increased VS-SPL and VS-MTG functional coupling for Test-FB-PE versus Test-PE. Asterisks indicate statistically significant effects: *$q$ < 0.05. aHC = left anterior hippocampus; VS = ventral striatum; MFG = middle frontal gyrus; IFG = inferior frontal gyrus; MTG = middle temporal gyrus; SPL = superior parietal lobe; SMG = supramarginal gyrus. Yellow edges in renderings = Analysis I, *Condition* main effect; orange edges in renderings = Analysis II, *Condition* main effect; cyan edges in renderings = Analysis II, *Performance in repetition cycle 1* main effect.

*Analysis II.* To investigate whether initial retrieval performance modulated functional coupling of the left anterior and posterior hippocampus, left SMG and left ventral striatum with the remaining regions of interest, word pairs tested with or without additional performance feedback were conditioned based on participant's retrieval performance in repetition cycle 1. A series of 2 x 2 repeated measures ANOVAs with the factors *Condition* (Test-PE vs. Test-FB-PE) and *Performance in repetition cycle 1* (correct vs. incorrect) were conducted. Anterior hippocampus, SMG and ventral striatum showed main effects of *Condition*, with SMG also showing main effects of Performance in repetition cycle 1, but importantly, no interactions were statistically significant. Results are reported in Table 2 (see also Figure 5 B, C & D). Furthermore, brain-behavior correlations revealed a positive relationship between final test retrieval and the increased functional coupling of left anterior hippocampus with left ventral striatum, but only for items previously tested with additional performance feedback that were retrieved incorrectly in repetition cycle 1 (Test-PE correct: $q$ = 0.740; Test-PE incorrect: $q$ = 0.685; Test-FB-PE correct: $q$ = 0.400; Test-FB-PE incorrect: $q$ = 0.048).

**Table 2: Summary of functional connectivity results for Analysis I and Analysis II**

**Analysis I (Restudy vs. Test-PE vs. Test-FB-PE)**

| | | Restudy vs. Test-PE | Test-PE vs. Test-FB-PE | Restudy vs. Test-FB-PE |
|---|---|---|---|---|
| **aHC-MTG** | $F(2, 52) = 3.67$, $q = .032$, $\eta_p^2 = .12$ | $q = .449$ | ***q* = .044** | ***q* = .044** |
| **aHC-SPL** | $F(2, 52) = 3.81$, $q = .032$, $\eta_p^2 = .13$ | $q = .628$ | ***q* = .045** | ***q* = .045** |
| **aHC-VS** | $F(2, 52) = 3.85$, $q = .032$, $\eta_p^2 = .13$ | $q = .403$ | ***q* = .030** | $q = .132$ |

**Analysis II (Test-PE vs. Test-FB-PE) X (correct vs. incorrect)**

| **Main effect of *Condition*** | |
|---|---|
| **aHC-MTG** | $F(1, 26) = 6.20$, $q = .039$, $\eta_p^2 = .19$ |
| **aHC-SPL** | $F(1, 26) = 10.14$, $q = .019$, $\eta_p^2 = .28$ |
| **aHC-VS** | $F(1, 26) = 11.47$, $q = .019$, $\eta_p^2 = .31$ |
| **aHC-SMG** | $F(1, 26) = 7.90$, $q = .031$, $\eta_p^2 = .23$ |
| **SMG-MTG** | $F(1, 26) = 5.14$, $q = .046$, $\eta_p^2 = .17$ |
| **SMG-SPL** | $F(1, 26) = 4.39$, $q = .046$, $\eta_p^2 = .14$ |
| **VS-MTG** | $F(1, 26) = 4.59$, $q = .046$, $\eta_p^2 = .15$ |
| **VS-SPL** | $F(1, 26) = 6.60$, $q = .039$, $\eta_p^2 = .20$ |
| **Main effect of *Performance in repetition cycle 1*** | |
| **SMG-MFG** | $F(1, 26) = 5.14$, $q = .046$, $\eta_p^2 = .17$ |
| **SMG-IFG** | $F(1, 26) = 4.63$, $q = .046$, $\eta_p^2 = .15$ |

## Discussion

The present study examined the neural correlates of TPE during successful memory retrieval. Specifically, we investigated whether explicit performance feedback after a retrieval attempt (1) further boosts performance relative to the classic TPE effect and (2) modulates the neural network associated with retrieving correct memory representations. Behaviorally, retrieval accuracy increased and retrieval speed decreased for the TPE compared to the restudy conditions. Importantly, additional performance feedback led to modest but significant further gains in retrieval performance. At the neural level, fronto-parietal brain regions, including MFG, IFG and SPL were more strongly engaged for restudied word pairs compared to those with retrieval attempts, whereas increased activation in temporo-parietal regions, in particular SMG and hippocampus, was observed for retrieval enhanced learning. Moreover, additional performance feedback did not affect regional activation, but instead significantly enhanced functional coupling of the ventral striatum as well as the SMG with hippocampus and temporo-parietal areas, including MTG and SPL.

Consistent with previous findings on the testing effect, our results confirmed an increase in retrieval accuracy and decreased RT for tested compared to restudied items in

the final memory test (see Roediger & Butler, 2011; Rowland, 2014; van den Broek, Segers, Takashima, & Verhoeven, 2014). Furthermore, additional performance feedback resulted in a 3 % increase in retrieval performance in the present study, consistent with our previous work (see Ludowicy et al., under review, showing also significant effects of additional performance feedback in two separate experiments with independent samples). This subtle but robust finding may have been moderated by the fact that the present paradigm was designed to induce high retrieval performance at the final test to obtain sufficient observations per condition for fMRI data analyses. Although no beneficial effect was detected in RTs, this could be due to the delay between the presentation of the cue and the response options in the covert retrieval test.

At present, few neuroimaging studies have investigated the testing effect specifically at the final retrieval test rather than at the repetition practice phase (Eriksson, Kalpouzos, & Nyberg, 2011; Keresztes et al., 2013; Rosner, et al., 2014; Wirebring et al., 2015; Wiklund-Hörnqvist et al., 2017, 2021). Those studies, which compared retrieval to restudy practice, have reported mixed findings, probably due to differences in the paradigms used to investigate the testing effect. Rosner et al. (2014) did not detect differences in neural activation at final test contrasting previously generated and restudied items, potentially due to the immediate administration of the final test or the indirect comparison of the test and restudy condition given they compared successful recognition versus correct rejection. In contrast, consistent with our findings, Keresztes et al. (2013) and Wiklund-Hörnqvist et al. (2017) reported an increase in neural activation in fronto-parietal networks for restudying compared to retrieval attempts.

On the one hand, the results of the present study revealed increased activation in fronto-parietal brain regions, especially in MFG, IFG and SPL for items previously restudied relative to those with retrieval practice. Hence, the present results support the notion that the two routes to encoding (i.e., TPE vs. restudying) differently affect successful memory retrieval at later stages, presumably by modulating cognitive control processes in frontal

cortices and facilitating semantic memory related processes in the parietal cortex (Keresztes et al., 2013; van den Broek et al., 2016; Wiklund-Hörnqvist et al., 2017).

On the other hand, our results revealed an increase in neural activation in temporo-parietal regions, including SMG, for correctly retrieved items previously practiced by testing as compared to restudying. Rosner et al. (2014) reported increased activation in several brain regions, including SMG, for successfully retrieved memories compared with correct rejections. Moreover, SMG has been implied in modulating the testing effect in prior literature, but so far only in studies focused on the practice phase instead of final retrieval. While increased activation in SMG for subsequently retrieved compared to forgotten items was reported during retrieval practice (van den Broek et al., 2013; Vannest et al., 2012), other research reported enhanced activation in SMG during restudy (Vestergren & Nyberg, 2014; Wing et al., 2013).

Both SMG and angular gyrus were previously proposed as relevant brain areas for the storage of semantic information (Binder, Desai, Graves, & Conant, 2009; van den Broek et al., 2016), but SMG has also been associated with the processing of phonological information (e.g., Oberhuber et al., 2016). The tasks used to investigate the testing effect sometimes require the processing of phonological details specific to the stimulus materials. For example, in the studies by Rosner et al. (2014) or Vannest et al. (2012), participants were asked to generate associated words to target words given only several letters as cues (e.g., *GARBAGE-W_ST_*), which requires the processing of phonological information to guess the associated word. In the present study, participants were asked to indicate the third letter of the target word during the repetition phase, which promotes complex phonological processing of the target word as well. Thus, testing could possibly promote later retrieval performance by systematically strengthening only relevant connections between certain details of the representations that were specifically needed during retrieval attempts, such as the phonological representations. Alternatively, the parietal cortex may play a critical role in episodic memory processing, with the ventral portion, including SMG, being a modulator of bottom-up attention or capturing attentional resources by relevant memory cues or recovered

memories [see attention to memory (AtoM) model by Cabeza, Ciaramelli, Olson, & Moscovitch, 2008]. Ventral parietal cortex is typically associated with recollection-related processes as well as higher mnemonic confidence levels (Cabeza et al., 2008), which are expected to be stronger for correctly retrieved items previously practiced by testing as compared to restudying.

The goal of the present study was to investigate the neural processes underlying the beneficial effect of performance feedback above and beyond the feedback that can be deduced from the correct answer during the restudy opportunity. Examining BOLD responses during the final retrieval test, we only detected changes in neural activation specific to items previously tested with or without additional performance feedback in the hippocampus. Prior neuroscientific studies on episodic memory highlighted the importance of hippocampal regions for retrieval enhanced learning, suggesting a complementary involvement with the default mode network (Jonker, Dimsdale-Zucker, Clarke, & Ranganath 2018). In line with this research, Wiklund-Hörnqvist et al. (2021) performed a testing effect study collecting fMRI data at the final retrieval test to examine the role of hippocampal subregions, reporting increased activation for tested compared to restudied items, especially in posterior hippocampus. In contrast to their findings, such activation patterns in anterior hippocampus were only observed in the present study when comparing items previously tested with additional performance feedback versus studied ones, but not for items tested without performance feedback. This difference might be due to the experimental designs, for example the retention interval between the repetition phase and the final test was one day in the present study compared to 7 days in the study by Wiklund-Hörnqvist et al. (2021), and materials were practiced three times in the present study compared to six times in the study by Wiklund-Hörnqvist et al. (2021), due to the difficulty of learning the material. Nevertheless, additional performance feedback seemed to boost hippocampal activation, leading to a similar activation pattern as reported in the study by Wiklund-Hörnqvist et al. (2021). Future studies might dissociate further factors influencing learning that result in increased hippocampal activation, as this could be related to consolidation.

In addition to examining regional activation levels in fMRI data, recent research has examined functional connectivity to investigate whether specific behaviors are associated with the functional coupling between brain regions (see Noble, Scheinost, & Constable, 2019). To the best of our knowledge, only one study has investigated the testing effect using functional connectivity analyses, focusing on testing effect-related coupling between hippocampus and other brain regions (Wing et al. 2013). These results revealed enhanced hippocampal coupling with posterior cingulate cortex, medial PFC and left IFG for tested compared to studied items (Wing et al. 2013). Accordingly, the present study investigated functional coupling between anterior and posterior hippocampus, ventral striatum and SMG with the remaining ROIs (MFG, IFG, MTG, SPL), revealing increased coupling between only the anterior portion of the left hippocampus with ventral striatum, SPL, SMG and MTG for items tested with additional performance feedback. Furthermore, increased functional connectivity was also observed for ventral striatum and SMG with MTG and SPL for items previously tested with additional performance feedback. In sum, the present results suggest that additional performance feedback during learning enhances later retrieval performance by strengthening the connectivity between the anterior hippocampus and temporo-parietal regions, as well as the ventral striatum and ventral parietal cortex connectivity with other temporo-parietal regions (MTG, SPL). In addition, brain-behavior correlations complemented this finding by revealing an increase in anterior hippocampus-ventral striatum functional coupling for items tested with additional performance feedback which were previously incorrectly retrieved in repetition cycle 1. These results highlight that the anterior hippocampus, ventral striatum and SMG act like hubs with their connectivity with temporo-parietal regions being associated with testing with correct answer feedback or additional performance feedback. Based on previous evidence, we think that this circuitry is related to reinforced memories via testing + feedback and that these results are consistent with the AtoM model (Cabeza et al, 2008), which proposes that ventral parietal cortex receives relevant bottom-up inputs from reinforced mnemonic representations. Also, importantly, the SMG does not seem to only act as a hub for the reinforced memories *via* additional

performance feedback, but also showed significant coupling with frontal regions (MFG, IFG) for items incorrectly versus correctly retrieved in repetition cycle 1, which may reflect its role in phonological processing and semantic elaboration, strengthening details of the representations that were specifically needed during retrieval attempts. Although future research is needed to further examine these two potential roles of SMG related to TPE with feedback as a function of its connections, our results suggest that this region may serve as an interface among available reinforced mnemonic inputs and further phonological/semantic elaboration processes needed to improve subsequent memory performance.

Recent literature aiming to explain the mechanisms underlying the testing effect has emphasized two theoretical accounts. On the one hand, the elaboration account (Carpenter & DeLosh, 2006; Carpenter & Olson, 2011; Pyc & Rawson, 2009) suggested that each testing situation helps to elaborate the semantic representation and hence increases the number of cues which support later successful retrieval. On the other hand, the search-set restriction account (e.g., Karpicke & Smith, 2012; Thomas & McDaniel, 2013) proposed that retrieval attempts strengthen correct associations by reducing the connections to competing information and hence refine the cue-related search-set. In line with these theories, we hypothesized that testing with additional performance feedback would cause increased activation in areas related to feedback processing, such as the ventral striatum, and areas related to semantic memory, such as the lateral temporal and parietal cortices. Our results revealed increased activation in frontal-parietal brain regions for previously studied compared to tested items, and only in temporal regions for tested compared to studied material. The present study supports the idea that testing reduces neural activation (van den Broek et al., 2016) in anterior PFC regions and helps to refine the search-set. Similarly, the results that additional performance feedback led to increased left hippocampal activation at the final test, but did not affect activation in ventral striatum might hint towards increased search set restrictions instead of elaboration of the semantic representations. However, functional connectivity results revealing increased coupling between ventral striatum, anterior hippocampus and SMG with temporo-parietal regions could support bottom-up attention to

process memory representations further enhanced via feedback. Moreover, functional coupling of the SMG with lateral PFC regions may support semantic elaboration and phonological processes further aiding memory retrieval.

The present study was aimed at investigating successful memory retrieval at the final test. Due to this focus, however, we were unable to differentiate whether the present findings result from modulation of either encoding or retrieval processes. Therefore, future studies might investigate encoding processes as well as both successful and unsuccessful memory retrieval. In addition, examining the repetition cycle instead of the final test may provide the possibility to capture performance feedback effects at TPE and to differentiate between the effects of positive and negative feedback. Potentially, other neuroscientific methods such as electroencephalography could provide additional information on the temporal processes underlying feedback-enhanced TPE.

In conclusion, the present study confirmed previous evidence by demonstrating that retrieval attempts as compared to restudying word pairs enhanced later memory retrieval (TPE). We extended this classic pattern of findings by also evaluating RTs, showing that better performance was also evident in a higher pace of retrieval. Critically, explicit performance feedback regarding the retrieval attempts provided just before the correct answer feedback further enhanced retrieval performance, and hence it might offer a useful tool in applied settings. Examining the neural underpinnings of successful retrieval of the target words one day after the repetition practice session allowed us to dissociate the influence of the different encoding conditions on the final retrieval processes: Stronger functional left lateral PFC engagement while retrieving items previously restudied versus items previously tested, and greater lateral temporo-parietal activation for retrieving items previously tested versus items previously restudied. Notably, performance feedback during the practice phase led to increased hippocampal activation at the final test on Day 2 and functional connectivity underscored the role of ventral striatum, hippocampus and SMG functional coupling for successfully retrieved memories tested with additional feedback. With respect to the mechanisms underlying the testing effect, these results provide support for the

search-set restriction account, the AtoM model and the elaboration account as well. These results add to the growing body of evidence suggesting that the testing effect is not due solely to quantitative changes that occur in memory performance, but also that qualitatively different networks are implied in retrieving correct memory traces encoded under different conditions.

## Authors' contributions

Conceptualization: P.L., D.C., and P.M.P.-A.; investigation: P.L., J.A.-T. and K.G.; formal analyses and methodology: P.L. and P.M.P.-A.; software: P.M.P.-A.; project administration, funding acquisition and validation: P.M.P.-A.; supervision: D.C., T.L. and P.M.P.A.; writing original draft: P.L. and P.M.P.-A.; writing review & editing: D.C., T.L. and K.G.


## Funding

P.L. was supported by the Rhineland-Palatinate Research Initiative (Potentialbereich Cognitive Dynamics) of the Federal Ministry of Science, Further Education and Culture (MWWK). This research was supported by funding from the Spanish Ministry of Science and Innovation (PID2021-123574NB-I00), Basque Government (PIBA-2021-1-0003), a grant from "la Caixa" Banking Foundation (under the project code LCF/PR/HR19/52160002PID2019-105520GB-100) to P.M.P-A. BCBL acknowledges funding from the Basque Government through the BERC 2022-2025 program and by the Spanish State Research Agency through BCBL Severo Ochoa excellence accreditation CEX2020-001010-S.

## Acknowledgements

The authors would like to thank Caroline Handley for proofreading and helpful comments on the manuscript.

Conflict of interest statement: None declared.



Correspondence regarding this manuscript should be addressed either to Petra Ludowicy, Center for Cognitive Science, University of Kaiserslautern, 67663, Kaiserslautern, Germany; or to Pedro M. Paz-Alonso, BCBL, Paseo Mikeletegi 69, 2, 20009 Donostia-San Sebastián, Spain. E-mails: pludowicy@gmail.com, kepa.pazalonso@gmail.com


## Data availability

Data are available at https://www.bcbl.eu/DataSharing/CerebCor2023Ludowicy/

## 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*(3), 940.

Binder, J. R., Desai, R. H., Graves, W. W., & Conant, L. L. (2009). Where is the semantic system? A critical review and meta-analysis of 120 functional neuroimaging studies. *Cerebral cortex*, *19*(12), 2767-2796.

Brett, M., Anton, J. L., Valabregue, R., & Poline, J. B. (2002, June). Region of interest analysis using an SPM toolbox. In *8th international conference on functional mapping of the human brain* (Vol. 16, No. 2, p. 497).

Cabeza, R., Ciaramelli, E., Olson, I. R., & Moscovitch, M. (2008). The parietal cortex and episodic memory: an attentional account. *Nature reviews neuroscience*, *9*(8), 613-625.

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*(2), 268-276.

Carpenter, S. K., & Olson, K. M. (2012). Are pictures good for learning new vocabulary in a foreign language? Only if you think they are not. *Journal of experimental psychology: learning, memory, and cognition*, *38*(1), 92.

Clos, M., Bunzeck, N., & Sommer, T. (2019). Dopamine enhances item novelty detection via hippocampal and associative recall via left lateral prefrontal cortex mechanisms. *Journal of Neuroscience*, *39*(40), 7920-7933.

Cocosco, C. A., Kollokian, V., Kwan, R. K. S., Pike, G. B., & Evans, A. C. (1997). Brainweb: Online interface to a 3D MRI simulated brain database. In *NeuroImage*.

Costa, V. D., Dal Monte, O., Lucas, D. R., Murray, E. A., & Averbeck, B. B. (2016). Amygdala and ventral striatum make distinct contributions to reinforcement learning. *Neuron*, *92*(2), 505-517.

Eriksson, J., Kalpouzos, G., & Nyberg, L. (2011). Rewiring the brain with repeated retrieval: a parametric fMRI study of the testing effect. *Neuroscience letters*, *505*(1), 36-40.

Ernst, B., & Steinhauser, M. (2012). Feedback-related brain activity predicts learning from feedback in multiple-choice testing. *Cognitive, Affective, & Behavioral Neuroscience*, *12*(2), 323-336.

Fischl, B., Sereno, M. I., & Dale, A. M. (1999). Cortical surface-based analysis: II: inflation, flattening, and a surface-based coordinate system. *Neuroimage*, *9*(2), 195-207.

Fisher, R. A. (1921). On the'probable error'of a coefficient of correlation deduced from a small sample. *Metron*, *1*, 1-32.

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

Jonsson, B., Wiklund-Hörnqvist, C., Stenlund, T., Andersson, M., & Nyberg, L. (2021). A learning method for all: The testing effect is independent of cognitive ability. *Journal of Educational Psychology*, *113*(5), 972.

Kahn, I., & Shohamy, D. (2013). Intrinsic connectivity between the hippocampus, nucleus accumbens, and ventral tegmental area in humans. *Hippocampus*, *23*(3), 187-192.

Karpicke, J. D., & Smith, M. A. (2012). Separate mnemonic effects of retrieval practice and elaborative encoding. *Journal of Memory and Language*, *67*(1), 17-29.

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*(11), 3025-3035.

Lansink, C. S., Goltstein, P. M., Lankelma, J. V., Joosten, R. N., McNaughton, B. L., & Pennartz, C. M. (2008). Preferential reactivation of motivationally relevant information in the ventral striatum. *Journal of Neuroscience*, *28*(25), 6372-6382.

Lisman, J. E., & Grace, A. A. (2005). The hippocampal-VTA loop: controlling the entry of information into long-term memory. *Neuron*, *46*(5), 703-713.

Liu, X. L., Liang, P., Li, K., & Reder, L. M. (2014). Uncovering the neural mechanisms underlying learning from tests. *PLoS One*, *9*(3), e92025.

Liu, X. L., & Reder, L. M. (2016). fMRI exploration of pedagogical benefits of repeated testing: when more is not always better. *Brain and behavior*, *6*(7), e00476.

Ludowicy, P., Paz-Alonso, K., Lachmann, T. & Czernochowski, D. (under review). Timing matters: immediate performance feedback enhances test-potentiated encoding.

Mazaika, P. K., Hoeft, F., Glover, G. H., & Reiss, A. L. (2009). Methods and software for fMRI analysis of clinical subjects. *Neuroimage*, *47*(Suppl 1), S58.

Nelson, S. M., Arnold, K. M., Gilmore, A. W., & McDermott, K. B. (2013). Neural signatures of test-potentiated learning in parietal cortex. *Journal of Neuroscience*, *33*(29), 11754-11762.

Nelson, D. L., McEvoy, C. L., & Schreiber, T. A. (2004). The University of South Florida free association, rhyme, and word fragment norms. *Behavior Research Methods, Instruments, & Computers*, *36*(3), 402-407.

Noble, S., Scheinost, D., & Constable, R. T. (2019). A decade of test-retest reliability of functional connectivity: A systematic review and meta-analysis. *Neuroimage*, *203*, 116157.

O'Doherty, J., Dayan, P., Schultz, J., Deichmann, R., Friston, K., & Dolan, R. J. (2004). Dissociable roles of ventral and dorsal striatum in instrumental conditioning. *science*, *304*(5669), 452-454.

Oberhuber, M., Hope, T. M. H., Seghier, M. L., Parker Jones, O., Prejawa, S., Green, D. W., & Price, C. J. (2016). Four functionally distinct regions in the left supramarginal gyrus support word processing. *Cerebral Cortex*, *26*(11), 4212-4226.

Olszowy, W., Aston, J., Rua, C., & Williams, G. B. (2019). Accurate autocorrelation modeling substantially improves fMRI reliability. *Nature communications*, *10*(1), 1-11.

Poppenk, J., & Moscovitch, M. (2011). A hippocampal marker of recollection memory ability among healthy young adults: contributions of posterior and anterior segments. *Neuron*, *72*(6), 931-937.

Poppenk, J., Evensmoen, H. R., Moscovitch, M., & Nadel, L. (2013). Long-axis specialization of the human hippocampus. *Trends in cognitive sciences*, *17*(5), 230-240.

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*(4), 437-447.

Rissman, J., Gazzaley, A., & D'Esposito, M. (2004). Measuring functional connectivity during distinct stages of a cognitive task. *Neuroimage*, *23*(2), 752-763.

Roediger III, H. L., & Butler, A. C. (2011). The critical role of retrieval practice in long-term retention. *Trends in cognitive sciences*, *15*(1), 20-27.

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

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

Rosner, Z. A., Elman, J. A., & Shimamura, A. P. (2013). The generation effect: Activating broad neural circuits during memory encoding. *cortex*, *49*(7), 1901-1909.

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

Scimeca, J. M., & Badre, D. (2012). Striatal contributions to declarative memory retrieval. *Neuron*, *75*(3), 380-392.

Spaniol, J., Davidson, P. S., Kim, A. S., Han, H., Moscovitch, M., & Grady, C. L. (2009). Event-related fMRI studies of episodic encoding and retrieval: meta-analyses using activation likelihood estimation. *Neuropsychologia*, *47*(8-9), 1765-1779.

Thomas, R. C., & McDaniel, M. A. (2013). Testing and feedback effects on front-end control over later retrieval. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *39*(2), 437.

Van den Broek, G., Takashima, A., Wiklund-Hörnqvist, C., Wirebring, L. K., Segers, E., Verhoeven, L., & Nyberg, L. (2016). Neurocognitive mechanisms of the “testing effect”: A review. *Trends in Neuroscience and Education*, *5*(2), 52-66.

van den Broek, G. S., Segers, E., Takashima, A., & Verhoeven, L. (2014). Do testing effects change over time? Insights from immediate and delayed retrieval speed. *Memory*, *22*(7), 803-812.

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

Vannest, J., Eaton, K. P., Henkel, D., Siegel, M., Tsevat, R. K., Allendorfer, J. B., ... & Szaflarski, J. P. (2012). Cortical correlates of self-generation in verbal paired associate learning. *Brain research*, *1437*, 104-114.

Vestergren, P., & Nyberg, L. (2014). Testing alters brain activity during subsequent restudy: evidence for test-potentiated encoding. *Trends in Neuroscience and Education*, *3*(2), 69-80.

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. *Journal of Neuroscience*, *35*(26), 9595-9602.

Wiklund-Hörnqvist, C., Andersson, M., Jonsson, B., & Nyberg, L. (2017). Neural activations associated with feedback and retrieval success. *npj Science of Learning*, *2*(1), 1-7.

Wiklund-Hörnqvist, C., Stillesjö, S., Andersson, M., Jonsson, B., & Nyberg, L. (2021). Retrieval practice facilitates learning by strengthening processing in both the anterior and posterior hippocampus. *Brain and behavior*, *11*(1), e01909.

Wing, E. A., Marsh, E. J., & Cabeza, R. (2013). Neural correlates of retrieval-based memory enhancement: an fMRI study of the testing effect. *Neuropsychologia*, *51*(12), 2360-2370.

**Supplementary Material**

Table S1

| Cluster | Local maxima | Hem | x | y | z | Z-score | Voxels (k) |
|---|---|---|---|---|---|---|---|
| Restudy > Test-PE / Test-FB-PE | | | | | | | |
| 1 | Pars triangularis | L | -45 | 20 | 29 | 6.89 | 1313 |
| | Pars orbitalis | L | -30 | 26 | 2 | 6.23 | |
| | Middle frontal gyrus | L | -51 | 14 | 41 | 5.77 | |
| | Precentral gyrus | L | -48 | 2 | 53 | 5.24 | |
| | Middle frontal gyrus | L | -35 | 8 | 62 | 5.00 | |
| | Pars orbitalis | L | -39 | 44 | -1 | 3.97 | |
| 2 | Inferior Temporal Gyrus | L | -42 | -52 | -7 | 5.32 | 280 |
| | Inferior Occipital Gyrus | L | -36 | -82 | -7 | 4.47 | |
| 3 | Cerebellum (Crus 2) | R | 12 | -79 | -34 | 5.48 | 272 |
| | Cerebellum (VIII) | R | 30 | -67 | -49 | 4.92 | |
| 4 | Middle Occipital Gyrus | L | -24 | -67 | 38 | 5.26 | 237 |
| | Superior Parietal Lobe | L | -27 | -58 | 44 | 4.84 | |
| study < test / testFB | | | | | | | |
| 1 | Superior temporal gyrus | L | -51 | -7 | 5 | 6.7 | 661 |
| | Superior temporal gyrus | L | -66 | -19 | 14 | 5.87 | |
| | Rolandic operculum | L | -60 | -22 | 14 | 5.86 | |
| | Rolandic operculum | L | -42 | -10 | 23 | 4.83 | |
| | Insula | L | -36 | -16 | 8 | 4.72 | |
| | Supramarginal gyrus | L | -60 | -31 | 26 | 3.89 | |
| 2 | Precentral gyrus | L | -27 | -31 | 62 | 6.53 | 458 |
| | Postcentral gyrus | L | -21 | -40 | 65 | 6.00 | |
| | Precuneus | L | -21 | -49 | 65 | 4.7 | |
| 3 | Rolandic operculum | R | 54 | -19 | 20 | 5.35 | 481 |
| | Temporal Pole | R | 66 | -1 | 5 | 5.18 | |
| | Superior temporal gyrus | R | 54 | -4 | 2 | 4.74 | |
| | Insula | R | 42 | -7 | -1 | 4.6 | |
| 4 | Precuneus | R | 12 | -43 | 62 | 4.89 | |
| | Postcentral | R | 27 | -40 | 71 | 4.77 | |
| | Precentral | R | 27 | -19 | 68 | 3.86 | |