Main Manuscript for

Neural and computational evidence for a predictive learning account of the testing effect

Haopeng Chen[1], Pieter Verbeke[2], Stefania Mattioni[1], Cristian Buc Calderon[3], Tom Verguts[1]

[1]Department of Experimental Psychology, Faculty of Psychology and Educational Sciences, Ghent University, Ghent, 9000, Belgium

[2]Howest University of Applied Sciences, Kortrijk, 8500, Belgium

[3]Centro Nacional de Inteligencia Artificial (CENIA), Santiago, 8320000, Chile

*Haopeng Chen.

Email: haopeng.chen@ugent.be

Author Contributions: H.C. designed the research, collected the data, performed data analysis, and wrote the manuscript. P.V. designed the research, collected the data, and performed data analysis. S.M. designed the research. C.B.C. designed the research and wrote the manuscript. T.V. designed the research and wrote the manuscript. All authors discussed the results and commented on the manuscript.

Competing Interest Statement: There is no competing interest.

Classification: Biological Sciences: Psychological and Cognitive Sciences.

Keywords: Testing effect; Predictive learning; fMRI; Ventral Striatum**.**

Abstract

Testing enhances memory more than studying. Although numerous studies have demonstrated the robustness of this classic effect, its neural and computational origin remains debated. Predictive learning is a potential mechanism behind this phenomenon: Because predictions and prediction errors (mismatch between predictions and feedback) can only be generated in testing (and not in studying), testing can benefit from predictive learning. We shed light on the testing effect from a multi-level analysis perspective via a combination of cognitive neuroscience experiments (fMRI) and computational modeling. At the neural level, we find that testing activates the canonical brain area related to reward prediction error, namely the ventral striatum. Crucially, activation in the ventral striatum fully mediates the testing effect. Computationally, only a model incorporating predictive learning can account for the full breadth of behavioral patterns observed in the data. These results provide strong and converging evidence for a predictive learning account of the testing effect.

Significance Statement

An exam is not a neutral measurement of memory. The *testing effect* entails that a test (e.g., an exam), is more effective than study for learning and memory. The same effect can be harnessed also before events of significance take place, rendering it an important aspect of an active learning strategy. Nevertheless, its origin remains unknown. We propose a novel predictive learning account, which posits that testing (but not studying) facilitates predictions about study material and promotes learning from prediction errors. Computationally, the testing effect was explained through a predictive-learning-based neural network. Neurally, the testing effect was mediated by the ventral striatum, a key brain region encoding prediction errors. This account may extend beyond testing to support active learning.

Main Text

Introduction

A remarkable finding is the testing effect, the robust phenomenon that testing enhances declarative memory retention more effectively than studying (1, 2). For instance, after an initial study session, taking tests instead of engaging in repeated study helps one to retain knowledge (3-5). Accordingly, learning apps such as Duolingo are increasingly advocating users to have tests and retrieval, rather than restudy. Previous work has attributed the testing effect to myriad factors, such as retrieval effort (6, 7) and elaborative semantic organization during testing (8, 9), which are absent during studying. Nevertheless, its underlying mechanisms remain a topic of debate. Here, we report and test an emergent predictive learning account as its cognitive and neural origin (10-12).

Predictive learning implies that minimizing prediction errors is a key objective for learning (13). This is a foundational principle in computational approaches to learning (including in Artificial Intelligence) and has more recently found its way into human declarative memory as well. Indeed, several studies have demonstrated that prediction errors can significantly promote declarative memory (14-16). For example, a student might initially predict that a dolphin is a type of fish, but is then corrected that it is, in fact, a mammal. Here, prediction errors can restructure and cement the student's information (neural representations) in memory.

Of importance for the present purpose, predictive learning can only appear in testing, but not in studying (10). Indeed, testing provides opportunities to predict possible answers, which are

absent in mere studying. Simply being informed that a dolphin is a mammal does not yield the same level of surprise relative to when one is forced to make a prediction beforehand. The mismatch between predictions and subsequent, external or internal, feedback generates prediction errors driving learning (10, 12). External feedback comes from the environment, while internal feedback can manifest as internal (confidence) feelings about whether a prediction is correct (17). Indeed, the neural signature of external feedback can also be active when people have an internal feeling of being correct or incorrect (18, 19). Importantly, both external and internal prediction errors are consistently localized in the midbrain across animal species including rodents (20), macaques (21, 22), and humans (17), where they are encoded by dopaminergic bursts. Recent studies suggest that the effect of prediction error on declarative memory can be fully mediated by the neural activation in the ventral striatum (23), a key region for dopamine release. Based on these results, we propose a novel dopaminergic neural basis (ventral striatum) for the testing effect and suggest that this neural basis can be explained from a predictive learning perspective. To substantiate this account, we employed a combination of cognitive neuroscience (fMRI) experiments and computational modeling techniques to investigate whether the testing effect activates and is mediated by ventral striatum activation, while evaluating the computational mechanisms of predictive learning giving rise to this effect.

Results

Experimental design.

Our declarative memory task consisted of four phases designed to help participants learn 90 Dutch-Swahili word pairs (Figure 1A) (10). In Phase 1, participants underwent initial learning, during which each word pair was displayed on the screen for 3 seconds. This phase ensured that participants acquired initial knowledge for the subsequent tests. Phase 2 involved a no-feedback assessment to control participants' initial learning performance. During this phase, participants selected the correct Swahili translation for a given Dutch word from four options and rated their confidence in their choice. Previous studies have used strategies such as excluding correct or high-confidence items from Phase 2 to filter out extreme initial learning (10). However, here and in previous studies (10), the behavioral patterns remain unchanged with or without applying these filters (See Supplementary Figure S1). For this reason, and to maintain strong statistical power in our fMRI analyses (Phase 3), we chose not to filter. The primary variable, Test vs. Study, was manipulated in Phase 3. Participants selected the correct Swahili translation of a Dutch word, either from four boxed options (classified as "test trials") or from a single boxed option containing the correct answer (classified as "study trials"). After making their selections, participants rated their confidence and received (informative but non-monetary) feedback with the correct answer. Phase 4 consisted of two final assessments without feedback. A word pair was considered fully learned only if it was correctly identified in both final assessments[1]. Only Phase 3 was conducted in the fMRI scanner.

Robust testing effect in behavioral data.

As shown in Figure 1B, the final assessment accuracy (Phase 4) after test is better than that after study, which replicates the robust testing effect ($\chi^2(1, N = 48) = 6.382$, $p = .012$, $\beta \pm 95\%\ CI = .036 \pm .029$, $R^2 = .001$). The testing effect became more robust ($\chi^2(1, N = 48) = 81.702$, $p < .001$,

[1] Final assessment accuracy was coded as 1 for correct responses in both final assessments, 0.5 for one correct response, and 0 for no correct response.

$\beta \pm 95\%\ CI$ = .225 ± .050, R^2 = .078) after controlling for feedback valence and confidence ratings from Phase 3 (Figure 1C-D).

The testing effect is only mediated by ventral striatum activation.

The fMRI analysis focused on neural activation during feedback onset. Contrasting beta maps (from Test-GLM; See Methods) of feedback onsets following test versus study revealed significant testing-specific activations (Figure 2) in the ventral striatum (VS), insula, superior frontal gyrus (SFG), inferior frontal gyrus (IFG), inferior parietal lobule (IPL), and midbrain.

Six mediating models were constructed to investigate the neural mechanisms underlying the testing effect, using the above-mentioned testing-related regions as neural mediators. The VS activation fully mediated the testing effect on final assessment accuracy (Figure 3A). In contrast, additional mediating models (Insula, SFG, IFG, IPL, midbrain) revealed no significant mediating effects (Figure 3B-F). Thus, consistent with our prediction, VS activation emerged as the primary neural basis of the testing effect.

Predictive learning model accounts for the behavioral data.

We next turned to computational modeling and constructed an associative memory neural network model (Figure 4A) where Dutch and Swahili words were presented as input and output, respectively (10). Initial connections between Dutch and Swahili units were based on participants' confidence ratings from Phases 2 and 3 (see Methods). The model then underwent testing and studying trials based on predictive learning, Hebbian learning, or a combination of both. Hebbian learning strengthens connections between neurons that are activated simultaneously (see Methods). Finally, the model completed a simulated final assessment to evaluate its performance. Based on different combinations of learning components (initial learning, predictive learning, and Hebbian learning), a total of seven models were constructed.

Model comparison (see Methods) suggested that the model with both initial learning and predictive learning (Model 5) fit the human data best (Figure 4B). Replicating previous research (10), model simulations showed that only the models incorporating predictive learning (Models 3, 5, 6, 7; Figure 4B) could replicate the testing effect in human data.

To examine the relationship between prediction errors and final assessment accuracy, we extracted trial-level prediction errors from the best-fitting model (Model 5). The influence of prediction errors on final assessment accuracy followed a W-shaped pattern (bottom right panel in Figure 4B), highlighting three key peaks. The first peak occurred at extremely negative prediction errors (around -1), and the second at extremely positive prediction errors (around 1), consistent with a critical role of (unsigned) prediction errors in memory consolidation (15, 24, 25). The third peak, at small positive prediction errors (around 0.1), was driven by initial learning rather than prediction errors. Specifically, strong initial learning enabled participants (and the model) to consistently provide correct answers in both Phase 3 and Phase 4, resulting in small prediction errors in Phase 3 but high final assessment accuracy in Phase 4. Indeed, the initial learning effect is obvious in the human behavioral pattern, in which participants' final assessment accuracy increases with correct confidence ratings in Phase 3 (Figure 1D). Importantly, this W-shaped relationship was consistent across all relevant datasets, including the behavioral data from the current fMRI study and two behavioral datasets with the same experimental design from a previous study (10) that we re-analyzed (see Supplementary Figure S2).

Prediction errors and Testing trigger overlapping VS activation.

Our theory also predicts an overlap between prediction error and testing in basal ganglia areas. Prediction error-related activations (from PE-GLM, see Methods) are shown in yellow in Figure

5A, overlapping substantially with testing-related activation in the VS and insula (green regions in Figure 5A). The overlapping regions in the VS and insula were extracted as regions of interest (ROIs). Both ROIs exhibited stronger responses following testing compared to studying (Figure 5B, C). However, their responses to prediction errors differed: VS activity increased with both positive and negative prediction errors in a signed manner (Figure 5D), whereas the insula encoded the absolute (unsigned) magnitudes of prediction errors (Figure 5E).

Prediction error effect is mediated by VS activation.

We next constructed mediating models to investigate the neural basis of the prediction error effect on final assessment accuracy. The overlapping regions associated with both prediction errors and Test (vs Study) during feedback onsets (VS and insula) were extracted as potential neural mediators. To account for its influence, feedback valence was included as a covariate in all mediating models. These analyses identified VS activation as a mediator between prediction errors and final assessment accuracy (Figure 5F). However, VS activation alone could not fully explain the prediction error effect. One potential reason for the absence of a complete mediating effect is the strongly nonlinear relationship between prediction error and final assessment accuracy, as previously demonstrated (bottom right panel of Figure 4B).

Discussion

This study combined behavioral, modeling, and fMRI approaches to explore predictive learning as a potential account for the testing effect. The behavioral and modeling results reveal prediction errors as the key driver of the testing effect. Furthermore, the fMRI results revealed that Test (vs Study) activates VS, an area strongly associated with prediction error-elicited dopamine release. Only VS activation could fully mediate the testing effect on final assessment accuracy, whereas other regions associated with testing (SFG, IFG, IPL, midbrain) could not. Moreover, prediction error-related and testing-related activation shared a common neural structure in VS.

VS activity as the neural basis of the testing effect.

The current study found that Test (vs. Study) elicited neural activations in several regions (VS, insula, SFG, IFG, IPL, and midbrain) during subsequent feedback. However, only VS activation fully mediated the relationship between testing and subsequent final assessment accuracy. Given that the VS is a key region for dopamine release and prediction error encoding (23, 26, 27), these findings suggest that prediction error may serve as the neural basis of the testing effect. In line with the current finding, some previous fMRI studies also found stronger activation in VS and semantic processing regions (IFG, IPL, middle temporal gyrus) during testing rather than studying (28). We speculate that the increased IFG, IPL, and middle temporal gyrus response during testing (relative to studying) may be modulated by VS activation during testing (relative to studying) (29), which could be explored by future studies.

Predictive learning as the origin of the testing effect.

Our computational modeling results indicate that the testing effect in human data could only be explained by an associative memory neural network model that incorporates predictive learning, and not Hebbian learning. This aligns with the computational principle that predictive learning is more efficient than Hebbian learning (30). Unlike Hebbian learning, predictive learning requires active retrieval or prediction, consistent with the cognitive and educational trends that advocate for active, prediction error-based learning (31, 32). Importantly, many previous theoretical accounts (extra retrieval effort, semantic organization, and feedback presentation during testing, rather than studying) of the testing effect are consistent with a predictive learning framework. First, retrieval is a critical process to generate prediction errors (6, 7). Second, an elaborative

semantic network can be organized through predictive learning (30, 33). Finally, the exact function of feedback might be to trigger prediction errors in a testing context.

The trial-level prediction errors estimated by the optimal predictive learning model influence final assessment accuracy in a W-shaped pattern. This pattern suggests that both extremely positive and negative prediction errors enhance declarative memory. Notably, extremely negative prediction errors are often triggered by high-confidence errors, which is labeled the hypercorrection effect (24, 34-36). The hypercorrection effect suggests that high-confidence errors are more easily corrected than low-confidence errors (36). Although several studies support the hypercorrection effect (24, 34, 35), some have reported contradictory findings, indicating that small errors are easier to correct (37). Considering our findings, one possible explanation is that studies failing to observe the hypercorrection effect might not elicit extremely negative prediction errors.

In our study, predictive learning was based on the contrast between external feedback and a prediction. However, the testing effect can occur in the absence of external feedback, leading to a classification of testing effects with external feedback (38, 39) and those without (3, 7). Although our predictive learning model accounts for the testing effect with external feedback, it requires modifications to explain the testing effect without external feedback. As mentioned earlier, numerous neural studies have shown that internal prediction error signals can emerge even when external feedback is unavailable (17, 28). We postulate that internal prediction errors may manifest as temporal prediction errors, resulting from discrepancies in predictions made at adjacent time points (13). For example, during a test, learners continuously accumulate evidence and refine their predictions. Initially, a learner might think that “mdudu” is the translation of the English word “wolf.” However, as one gathers more evidence, one may confidently determine that another Swahili word, "atay," is the correct answer. One interesting avenue for future research is evaluating how shifts in confidence and predictions, could generate temporal prediction errors that may influence memory encoding.

Prediction errors and testing share VS activity.

Like testing, our results revealed that also prediction errors robustly activate VS and insula. Importantly, only the overlapping (between testing and prediction errors) VS activity mediated the prediction error effect on the final assessment accuracy. Moreover, consistent with previous findings, VS and insula encode prediction errors in distinct ways (23, 40). Specifically, VS activation encodes the original prediction errors as signed prediction errors (23), while the insula encodes the absolute magnitude of the prediction errors as unsigned prediction errors (40). This finding may provide a crucial piece of the puzzle with respect to the debate of whether declarative memory is influenced by signed versus unsigned prediction errors (41). Indeed, future studies could explore how VS and insula leave their respective signed or unsigned prediction error signatures on declarative memory. Additionally, a surprising result in our work is that VS responses to positive prediction errors are not stronger than to negative prediction errors. In contrast, previous studies implementing monetary reward found that the VS response to positive (reward) prediction errors is stronger than to negative (reward) prediction errors (23). This suggests that monetary reward is important to regulate the VS encoding of signed prediction errors.

Dopaminergic signals are broadly transmitted across brain regions, including VS, hippocampus, and neocortex (27, 42), forming a distributed learning system (23, 29). Indeed, previous studies have reported VS activation and its functional connectivity with neocortex when participants successfully acquire the meaning of a novel word (29). Additionally, VS-hippocampus connectivity emerges during Swahili word learning and is modulated by reward prediction errors (23), suggesting that prediction errors during testing play a crucial role in activating the above complementary learning system. Interestingly, prediction errors may arise from hippocampus-

neocortex interactions (11, 12, 43). In particular, the hippocampus may gradually reconstruct incomplete information into an intact representation and use it as feedback for neocortex, potentially explaining how individuals generate (temporally) varying predictions and temporal prediction errors during testing (33, 44).

Summary

In summary, the current study supports the notion that the testing effect originates from predictive learning, as evidenced by cognitive neuroscience and modeling findings. Notably, the testing effect represents just one instance of a broader range of active learning strategies that emphasize active predictions (45, 46). Predictive learning may not only serve as a specific mechanism underlying the testing effect but also offer a broader cognitive framework for general active learning approaches, such as generation effect (45), problem-based learning (31, 32) , and error-driven learning (24, 46).

Materials and Methods

Participants

Fifty-five participants, all native Dutch speakers, with no reports of smoking, alcohol consumption, medication use within 24 hours of scanning, or psychiatric disorders, were recruited for this study. Each participant was compensated 35€. Data from seven participants were excluded from the final analysis: six due to technical issues during the experiment, and one due to being left-handed. This resulted in a final sample of 48 right-handed participants (mean age = 22.44, SD = 2.65, range = 18–29; 34 female participants, 14 male participants). The study was approved by the local ethics committee at Ghent University Hospital, and all participants provided informed consent.

Experimental design

Participants were instructed to learn 90 Dutch-Swahili word pairs to the best of their ability (Figure 1A). They underwent 4 phases: Initial learning (Phase 1), Initial assessment (Phase 2), Test vs Study (Phase 3), and Final assessments (Phase 4). Given our primary interest in the neural mechanism underlying Test vs Study comparison, only Phase 3 was conducted in the fMRI scanner.

Phase 1: Initial learning

Participants were required to familiarize themselves with all word pairs before proceeding to the subsequent phases. All word pairs were randomly presented on the screen, with each pair displayed for 3 seconds.

Phase 2: Test without feedback

Following the initial learning, a test without feedback was conducted in Phase 2 to assess participants' learning progress before the formal manipulations in Phase 3. In each trial, participants selected the Swahili translation of a randomly presented Dutch word from four random options by pressing "f", "v", "n", or "j" on the keyboard, followed by a 4-point confidence rating ("very uncertain," "uncertain," "certain," or "very certain") for their choice. The resulting test accuracies and confidence ratings were used to evaluate their initial learning.

Phase 3: Test vs Study

Phase 3 was the critical phase, designed to manipulate the Test vs Study conditions and induce prediction errors. In each trial, participants were presented with a randomly selected Dutch cue and four random Swahili options, some of which were boxed. Participants could only choose from the boxed options, provide a confidence rating, and then receive feedback indicating whether their choice was correct ("Correct" in green or "Wrong" in red), along with the correct answer. Responses were made using a Cedrus Lumina LS-Pair MRI-compatible ergonomic response pad, where participants used their left middle finger, left index finger, right index finger, and right middle finger to select corresponding options. In testing trials, participants selected from four framed Swahili options, allowing them to predict the possible answer and experience the prediction error when feedback was provided. In contrast, during studying trials, the correct Swahili option was directly boxed, leaving no opportunity for prediction errors as participants could only select the correct answer. Testing accuracy and confidence ratings during this phase were also used as measures of initial learning. Additionally, three fixation intervals with optimized jittering were inserted before the onsets of the question (Jitter 1), confidence rating (Jitter 2), and feedback (Jitter 3). Since feedback onset was the most critical, Jitters 1 and 3 were longer than Jitter 2. Specifically, Jitters 1 and 3 were sampled from a Poisson distribution with a mean of 4.35 seconds (ranging from 3 to 8 seconds), while Jitter 2 was sampled from a shorter Poisson distribution with a mean of 2.35 seconds (ranging from 1 to 6 seconds).

Phase 4: Final assessments

Two final assessments were conducted to evaluate the final learning performance, following the same procedure as in Phase 2. Administering two assessments was diagnostic to know whether participants selected the correct answer based on actual knowledge rather than chance. Specifically, a Swahili answer was considered fully learned only if a participant selected this correct answer in both assessments, which is unlikely to occur by chance on two consecutive assessments.

Behavioral data analysis

A Generalized Linear Mixed Model (GLMM) was employed to analyze the trial-level data. The predictors in the model included Test vs. Study (Test = 1, Study = 0), Feedback (Positive = 1, Negative = 0), and confidence ratings (0, 0.33, 0.67, 1). The model incorporated subject-specific random slopes for predictors and a random intercept[2]. Final assessment accuracy was the dependent variable, coded as 1 for correct responses in both final assessments (Phase 4 and Phase 5), 0.5 for one correct response[3], and 0 for no correct response[4].

Modelling analysis

Model architecture

Several neural networks incorporating Hebbian learning and/or predictive learning were developed to fit the behavioral data (Figure 4A). By adjusting the combination of learning components within the model, we assessed which components were critical for replicating the testing effect. For instance, if the testing effect could only be replicated when predictive learning was included, it suggests that predictive learning is essential for the testing effect. Besides, if the predictive learning model is the best-fitting model, its optimized parameters would be used to

---

2. Statistical results remained unchanged when only a random intercept was added.
3. A score of 0.5 means a lucky guess rather than intermediate knowledge between know and not know.
4. Another binary strategy encoded final assessment accuracy as 1 for correct responses in both final assessments, and 0 otherwise, which led to the same statistical results.

estimate trial-level prediction errors in human data, which would then be applied in the fMRI analysis to investigate the neural basis of prediction errors.

The model's task was to establish accurate connections between 90 Dutch units in the input layer and 360 Swahili units in the output layer. Its initial connections were aligned with participants' confidence ratings from Phase 2 and Phase 3, simulating the initial learning of participants from Phase 1 (10). The exact equation is shown below:

$$w_{ij}^{0} = \alpha \times \left(c_{ij}^{2} + c_{ij}^{3}\right)$$

In this equation, c_{ij}^{2} and c_{ij}^{3} represent the confidence ratings from Phase 2 and 3, respectively. The relationship between confidence ratings and initial connections was scaled by parameter α. In the study condition, participants were given the correct answer directly and could not express their internal confidence in Phase 3, meaning we had no confidence data for those items in Phase 3. To address this, we substituted study-trial confidence c_s with the mean confidence rating from the testing condition:

$$c_s = \frac{\sum_i^I \sum_j^J c_{ij}^{t}}{N_t}$$

for each studied (rather than tested) word.

After establishing the initial connections, the model proceeded with testing and studying trials, corresponding to Phase 3 in the human task. In the testing trials, the model selected the correct Swahili translation of a Dutch word from four Swahili options, while in the studying trials, the model could only choose the correct Swahili answer. The connections between Dutch and Swahili words were updated using Hebbian learning or predictive learning (or both). Hebbian learning strengthened the connections between Dutch and Swahili targets in reward contexts (correct testing and studying trials), while predictive learning adjusted the connections based on prediction errors in testing trials (regardless of correctness), allowing for the separation of the contributions of the two learning components.

Regardless of which learning component was applied, the model generated its predictions according to the following equation:

$$\hat{y}_j = \frac{1}{1 + e^{-s\left(\sum_i^I w_{ij} x_i - b\right)}}$$

Thus, the activation of each element $\hat{y}_j$ in the output layer was determined by the weighted (w_{ij}) sum of the input values x_i, which was then passed through the sigmoid function in the above equation with inverse temperature s. Additionally, in the output layer, elements that were not options in the current trial were set as 0.

Hebbian learning, modulated by reward, was implemented as follows:

$$\Delta w_{ij} = \beta \times x_i \times y_j \times r_j$$

In this equation, the model updated the weight w_{ij} only when a reward r_j was present (i.e., when feedback was positive). Note that Hebbian learning not modulated by reward would promote learning in testing and studying conditions equally, thus was impossible to explain the testing effect.

Predictive learning followed the standard Rescorla-Wagner learning model (47):

$$\Delta w_{ij} = \beta \times x_i \times \left(y_j - \hat{y}_j\right)$$

Here, the model updated the weight w_{ij} based on the error between the actual feedback y_j and the model's prediction $\hat{y}_j$.

After the learning phase, the model completed a final assessment to evaluate its learning performance, replicating the procedure used in Phase 4 of the human task. The activations in the output layer were normalized as follows:

$$p_j = \frac{\hat{y}_j}{\sum_n^N \hat{y}_n}$$

with the activation level (p_j) of the correct Swahili unit serving as an indicator of the model's accuracy for each trial.

Model fitting

Since the main goal was to replicate human responses in Phase 4, the model's final assessment results were used to fit the human final assessment data, which helped determine the optimal model parameters. The optimization was conducted by maximizing the log-likelihood function for each participant's data. Parameters α and β were allowed to range from 0 to 1, while s and b were allowed to range from 0 to 10. Additionally, the sum of α and β was constrained to 1, ensuring the relative contribution of initial learning versus Hebbian learning (and/or predictive learning) was accounted for.

Model comparison

Importantly, we developed seven models based on all possible combinations of the three learning components mentioned earlier ($2^3 - 1 = 7$): initial learning, Hebbian learning, and predictive learning. The weighted Akaike information criterion (wAIC) was employed to evaluate and compare the model fits (48).

fMRI data acquisition

fMRI data were collected with a 3T Prisma scanner system (Siemens), with a 64-channel head coil. A 3D high-resolution anatomic image was obtained using a T1-weighted MPRAGE sequence [flip angle (FA) = 9°; matrix size = 176 × 256 × 256; interpolated voxel size = 1.0 × 1.0 × 1.0mm]. In addition, we acquired a field map per participant, to correct for magnetic field inhomogeneities (TR= 528ms; TE = 7.38ms; FA= 60°). Whole-brain functional images were acquired using a T2p-weighted EPI sequence (TR = 1780ms; TE = 27ms; FA= 66°; matrix size = 54 × 64 × 64; voxel size = 2.5 × 2.5 × 2.5 mm).

fMRI data analysis

Preprocessing

The fMRI data analysis was conducted using SPM 12 and MATLAB. After converting the raw DICOM data to BIDS format, the first five functional images for each participant were discarded to avoid transient spin saturation effects. The remaining functional images underwent slice timing correction, realignment, normalization, and smoothing. The structural image was co-registered with the mean realigned functional image. Task-related regressors were convolved with the

canonical Hemodynamic Response Function (HRF) in the General Linear Model (GLM) to model the BOLD signal.

Whole brain analysis

Two GLMs were constructed to examine neural correlates of Test vs. Study (during feedback onset) and prediction errors. In both models, the onsets for fixation, test (or study), confidence, and feedback were modeled as delta functions. Since prediction errors occur at feedback onset, the analysis primarily focused on neural activity during this stage. The first GLM (Test-GLM) included three main regressors for feedback onsets: feedback after correct testing, feedback after incorrect testing, and feedback after studying. Beta maps from these regressors were used to create individual Test vs. Study contrasts for group-level analysis, where significantly positive activations (False Discovery rate; FDR $p < 0.05$) indicated stronger neural responses to feedback after testing compared to studying. The second GLM (PE-GLM) modeled feedback onsets as a single regressor, incorporating trial-level mean-centered prediction errors (from the best-fitting computational model) and feedback valence as parametric modulators. Beta maps for the prediction error modulator were used in group-level analyses, with significantly positive activations (FDR $p < 0.05$) reflecting neural correlates of prediction error magnitude.

Mediating models

Brain regions associated with Test vs. Study contrast and prediction errors were extracted as regions of interest (ROIs) for mediating analysis. Two types of mediating models were constructed using conditional-level beta values (correct test, incorrect test, and study of each participant) in these ROIs as mediators and final assessment accuracy as the dependent variable. The first type of model examined the neural mediator of the testing effect, with Test vs. Study as the predictor. The second type of model focused on the neural mediator of the prediction error effect, with prediction errors (mean prediction errors in correct test, incorrect test, and study of each participant) as the predictor.

Acknowledgments

We thank Jonas Simoens, Ruth Krebs, and Filip Van Opstal for useful discussions about the topic of this study. H.C. received support from China Scholarship Council [CSC202206990008]. C.B.C. discloses support for the publication of this work from Centro Nacional de Inteligencia Artificial [FB210017]. The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript.

Code and data availability statement

All data, analysis code, and research materials are available at https://github.com/CogComNeuroSci/Haopeng_Public.

References

1. K. B. McDermott, Practicing Retrieval Facilitates Learning. *Annual Review of Psychology* **72**, 609-633 (2021).
2. H. L. Roediger, A. C. Butler, The critical role of retrieval practice in long-term retention. *Trends in Cognitive Sciences* **15**, 20-27 (2011).
3. H. L. Roediger, J. D. Karpicke, Test-enhanced learning: Taking memory tests improves long-term retention. *Psychological science* **17**, 249-255 (2006).
4. B. Jonsson, C. Wiklund-Hörnqvist, T. Stenlund, M. Andersson, L. Nyberg, A learning method for all: The testing effect is independent of cognitive ability. *Journal of Educational Psychology* **113**, 972-985 (2021).
5. T. R. Jonker, H. Dimsdale-Zucker, M. Ritchey, A. Clarke, C. Ranganath, Neural reactivation in parietal cortex enhances memory for episodically linked information. *Proceedings of the National Academy of Sciences* **115**, 11084-11089 (2018).
6. J. D. Karpicke, H. L. Roediger, Expanding retrieval practice promotes short-term retention, but equally spaced retrieval enhances long-term retention. *Journal of Experimental Psychology: Learning, Memory, and Cognition* **33**, 704-719 (2007).
7. M. A. Pyc, K. A. Rawson, 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 (2009).
8. S. K. Carpenter, E. L. DeLosh, Impoverished cue support enhances subsequent retention: Support for the elaborative retrieval explanation of the testing effect. *Memory & cognition* **34**, 268-276 (2006).
9. S. K. Carpenter, 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 (2009).
10. H. Chen, C. Hauspie, K. Ergo, C. Buc Calderon, T. Verguts, Predictive learning as the basis of the testing effect. *Communications Psychology* **3**, 18 (2025).
11. Y. Zheng, X. L. Liu, S. Nishiyama, C. Ranganath, R. C. O'Reilly, Correcting the hebbian mistake: Toward a fully error-driven hippocampus. *PLOS Computational Biology* **18**, e1010589 (2022).
12. X. L. Liu, R. C. O'Reilly, C. Ranganath, "Chapter Four - Effects of retrieval practice on tested and untested information: Cortico-hippocampal interactions and error-driven learning" in Psychology of Learning and Motivation, K. D. Federmeier, L. Sahakyan, Eds. (Academic Press, 2021), vol. 75, pp. 125-155.
13. R. Sutton, A. Barto, *Reinforcement learning: an introduction* (MIT, Cambridge, MA, 2018).
14. E. De Loof *et al.*, Signed reward prediction errors drive declarative learning. *PLoS One* **13**, e0189212 (2018).
15. N. Rouhani, K. A. Norman, Y. Niv, Dissociable effects of surprising rewards on learning and memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition* **44**, 1430 (2018).
16. A. Greve, E. Cooper, A. Kaula, M. C. Anderson, R. Henson, Does prediction error drive one-shot declarative learning? *Journal of Memory and Language* **94**, 149-165 (2017).
17. R. Daniel, S. Pollmann, Striatal activations signal prediction errors on confidence in the absence of external feedback. *NeuroImage* **59**, 3457-3467 (2012).

18. T. D. Satterthwaite *et al.*, Being right is its own reward: Load and performance related ventral striatum activation to correct responses during a working memory task in youth. *NeuroImage* **61**, 723-729 (2012).
19. C. B. Holroyd *et al.*, Dorsal anterior cingulate cortex shows fMRI response to internal and external error signals. *Nature neuroscience* **7**, 497-498 (2004).
20. N. Eshel, J. Tian, M. Bukwich, N. Uchida, Dopamine neurons share common response function for reward prediction error. *Nature Neuroscience* **19**, 479-486 (2016).
21. P. Montague, P. Dayan, T. Sejnowski, A framework for mesencephalic dopamine systems based on predictive Hebbian learning. *The Journal of Neuroscience* **16**, 1936-1947 (1996).
22. W. Schultz, P. Dayan, P. R. Montague, A Neural Substrate of Prediction and Reward. *Science* **275**, 1593-1599 (1997).
23. C. B. Calderon *et al.*, Signed Reward Prediction Errors in the Ventral Striatum Drive Episodic Memory. *J Neurosci* **41**, 1716-1726 (2021).
24. L. K. Fazio, E. J. Marsh, Correcting false memories. *Psychological science* **21**, 801-803 (2010).
25. J. Metcalfe, B. Finn, People's hypercorrection of high-confidence errors: Did they know it all along? *Journal of Experimental Psychology: Learning, Memory, and Cognition* **37**, 437 (2011).
26. R. de la Fuente-Fernández *et al.*, Dopamine release in human ventral striatum and expectation of reward. *Behavioural Brain Research* **136**, 359-363 (2002).
27. M. Watabe-Uchida, N. Eshel, N. Uchida, Neural Circuitry of Reward Prediction Error. *Annu Rev Neurosci* **40**, 373-394 (2017).
28. G. S. E. van den Broek, A. Takashima, E. Segers, G. Fernández, L. Verhoeven, Neural correlates of testing effects in vocabulary learning. *NeuroImage* **78**, 94-102 (2013).
29. P. Ripollés *et al.*, The Role of Reward in Word Learning and Its Implications for Language Acquisition. *Current Biology* **24**, 2606-2611 (2014).
30. T. Verguts, *Introduction to Modeling Cognitive Processes* (MIT Press, 2022).
31. D. F. Wood, Problem based learning. *BMJ* **326**, 328-330 (2003).
32. M. Prince, Does Active Learning Work? A Review of the Research. *Journal of Engineering Education* **93**, 223-231 (2004).
33. J. L. McClelland, B. L. McNaughton, R. C. O'Reilly, Why there are complementary learning systems in the hippocampus and neocortex: insights from the successes and failures of connectionist models of learning and memory. *Psychol Rev* **102**, 419-457 (1995).
34. B. Butterfield, J. A. Mangels, Neural correlates of error detection and correction in a semantic retrieval task. *Cognitive Brain Research* **17**, 793-817 (2003).
35. J. Metcalfe, B. Butterfield, C. Habeck, Y. Stern, Neural correlates of people's hypercorrection of their false beliefs. *Journal of Cognitive Neuroscience* **24**, 1571-1583 (2012).
36. B. Butterfield, J. Metcalfe, Errors committed with high confidence are hypercorrected. *Journal of Experimental Psychology: Learning, Memory, and Cognition* **27**, 1491 (2001).
37. T. Seabrooke, C. J. Mitchell, A. J. Wills, A. B. Inkster, T. J. Hollins, The benefits of impossible tests: Assessing the role of error-correction in the pretesting effect. *Memory & Cognition* **50**, 296-311 (2022).

38. N. Kornell, M. J. Hays, R. A. Bjork, Unsuccessful retrieval attempts enhance subsequent learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition* **35**, 989-998 (2009).
39. N. Kornell, P. J. Klein, K. A. Rawson, Retrieval attempts enhance learning, but retrieval success (versus failure) does not matter. *Journal of Experimental Psychology: Learning, Memory, and Cognition* **41**, 283 (2015).
40. C. W. Hoy *et al.*, Asymmetric coding of reward prediction errors in human insula and dorsomedial prefrontal cortex. *Nature Communications* **14**, 8520 (2023).
41. K. Ergo, E. De Loof, T. Verguts, Reward Prediction Error and Declarative Memory. *Trends in Cognitive Sciences* **24**, 388-397 (2020).
42. S. J. Sara, The locus coeruleus and noradrenergic modulation of cognition. *Nature Reviews Neuroscience* **10**, 211-223 (2009).
43. W. Sun, M. Advani, N. Spruston, A. Saxe, J. E. Fitzgerald, Organizing memories for generalization in complementary learning systems. *Nature Neuroscience* **26**, 1438-1448 (2023).
44. M. Meeter, J. M. J. Murre, Tracelink: A model of consolidation and amnesia. *Cognitive Neuropsychology* **22**, 559-587 (2005).
45. S. Bertsch, B. J. Pesta, R. Wiscott, M. A. McDaniel, The generation effect: A meta-analytic review. *Memory & Cognition* **35**, 201-210 (2007).
46. S. S. H. Wong, S. W. H. Lim, The derring effect: Deliberate errors enhance learning. *Journal of Experimental Psychology: General* **151**, 25-40 (2022).
47. R. A. Rescorla, A. R. Wagner, "A theory of Pavlovian conditioning: variations in the effectiveness of reinforcement and nonreinforcement" in Classical Conditioning II: Current Research and Theory, A. H. Blake, W. F. Prokasy, Eds. (Appleton-Century-Croft, 1972), pp. 64–99.
48. E.-J. Wagenmakers, S. Farrell, AIC model selection using Akaike weights. *Psychonomic bulletin & review* **11**, 192-196 (2004).

Figures and Tables

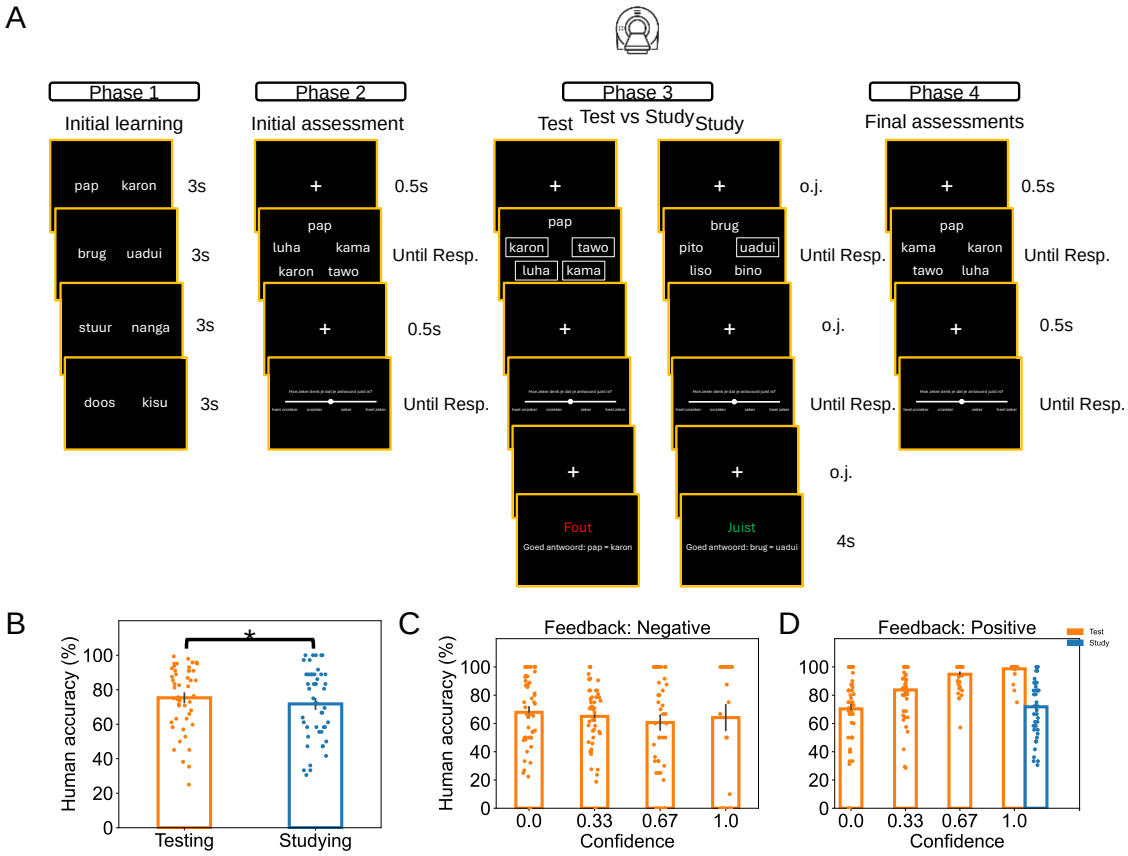

Figure 1. Experimental procedure (A) and behavioral results (B-D). A. Participants completed a four-phase task to learn 90 English-Swahili word pairs. Phase 1: All word pairs were presented randomly, with each pair displayed for 3 seconds to facilitate initial learning. Phase 2: An assessment without feedback was administered to assess initial learning. In each trial, participants selected the Swahili translation of an English cue from four options and rated their confidence in being correct. Phase 3: This phase introduced the Test vs. Study manipulation and prediction errors. In testing trials, participants selected the Swahili translation from four boxed options, while in studying trials, they could only choose the boxed correct answer. After each choice, they rated their confidence and received feedback, including whether they were correct and the correct answer. Importantly, three optimized jitter intervals were inserted between the choice onsets, confidence rating, and feedback to ensure precise timing and minimize expectancy effects. Phase 4: Two final assessments, using the same procedure as Phase 2 (test without feedback), were conducted to evaluate final learning performance. **B.** Behavioral pattern in the current study replicated the robust testing effect. **C-D.** The testing effect was still robust after controlling feedback valance and confidence ratings from Phase 3. Individual data points are represented by the dots; The error bars reflect the standard errors; Confidence ratings were from Phase 3; $p < 0.05$*.

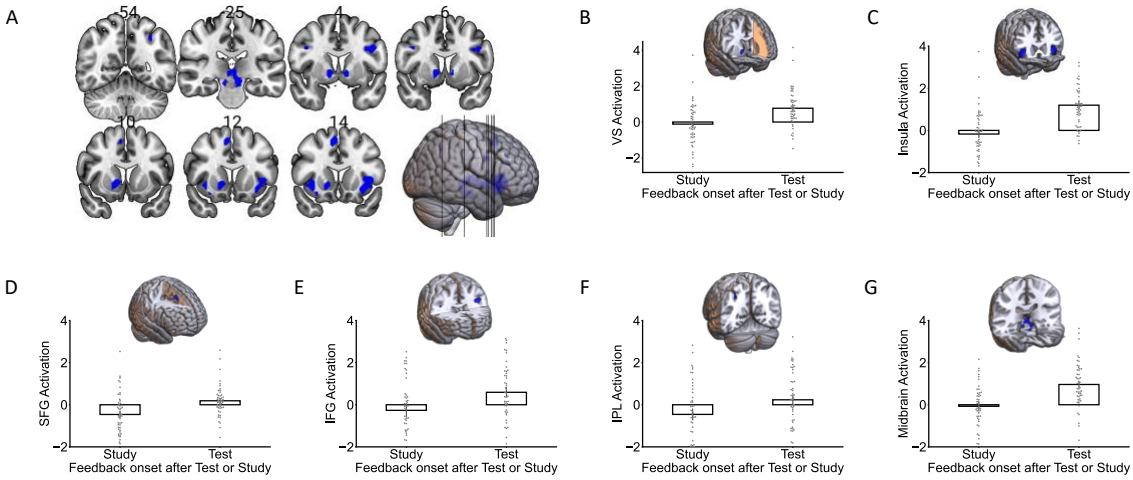

Figure 2. Test (vs Study) elicited neural activations during feedback onsets. A: Test mainly triggered neural activations in the VS, insula, SFG, IFG, IPL, and midbrain during feedback onsets. **B-G:** The mean neural activations across voxels in the testing-related VS, insula, SFG, IFG, IPL, and midbrain regions during feedback onsets.

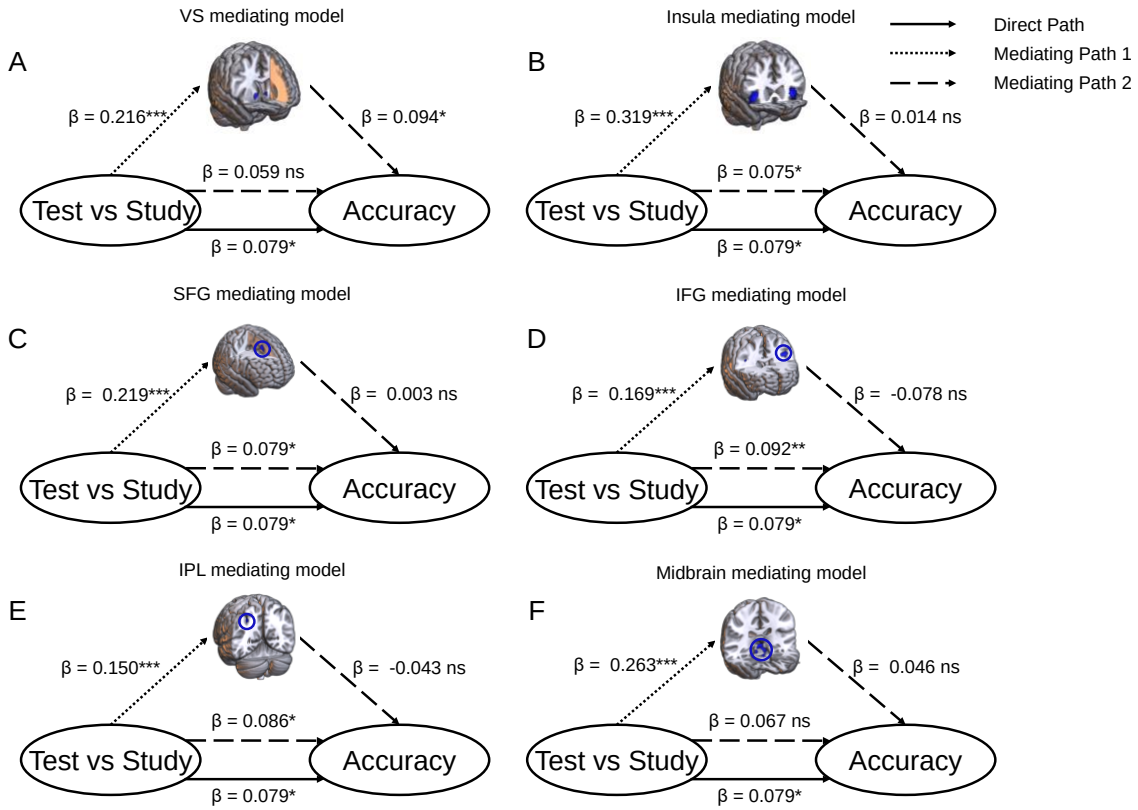

Figure 3. Neural mediation models for the testing effect. A: VS activation could fully mediate the testing effect on final assessment accuracy. **B-F:** Insula, SFG, IFG, IPL, and midbrain activations could not mediate the testing effect on final assessment accuracy. $p < 0.05^{*}$, $p < 0.01^{**}$, $p < 0.001^{***}$, $p > 0.05$ ns.

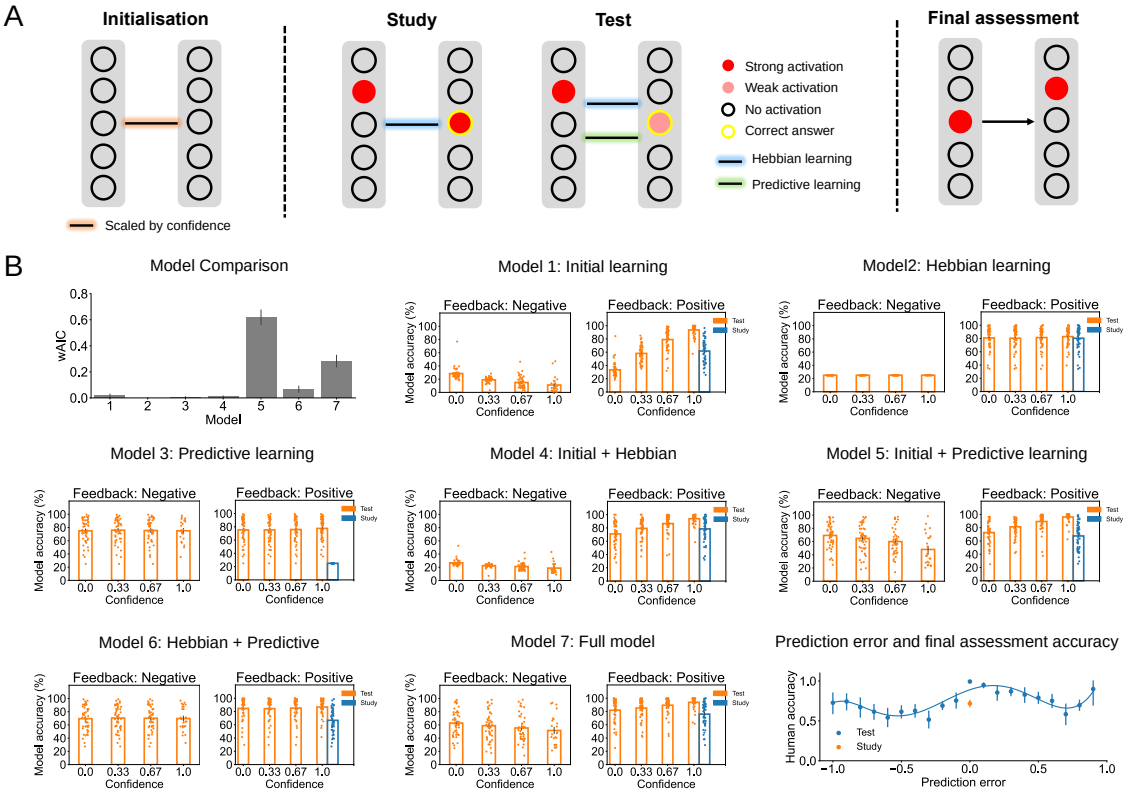

Figure 4. Model architecture (A) and modeling results (B). A. An associative memory neural network with 90 Dutch input units and 360 Swahili output units was developed to simulate human learning behavior. The model's initial connections were scaled according to human confidence ratings from Phase 2 and Phase 3. After initial connections, the model went through testing and studying trials as in Phase 3 of the human task. The model updated connections via Hebbian learning during both studying and testing trials. Predictive learning, however, was applied exclusively during testing trials. Final assessment was conducted as in Phase 4 of the human task. **B.** The model comparison and model simulations suggested that the model (Model 5) with both initial learning and predictive learning was the best model. Moreover, only models with predictive learning (Models 3, 5, 6, 7) could replicate the testing effect. Bottom right panel: Both extremely positive (around 1) and negative prediction errors (around -1) enhance the final assessment accuracy. The peak point around small positive prediction errors (around 0.1) was triggered by good initial learning.

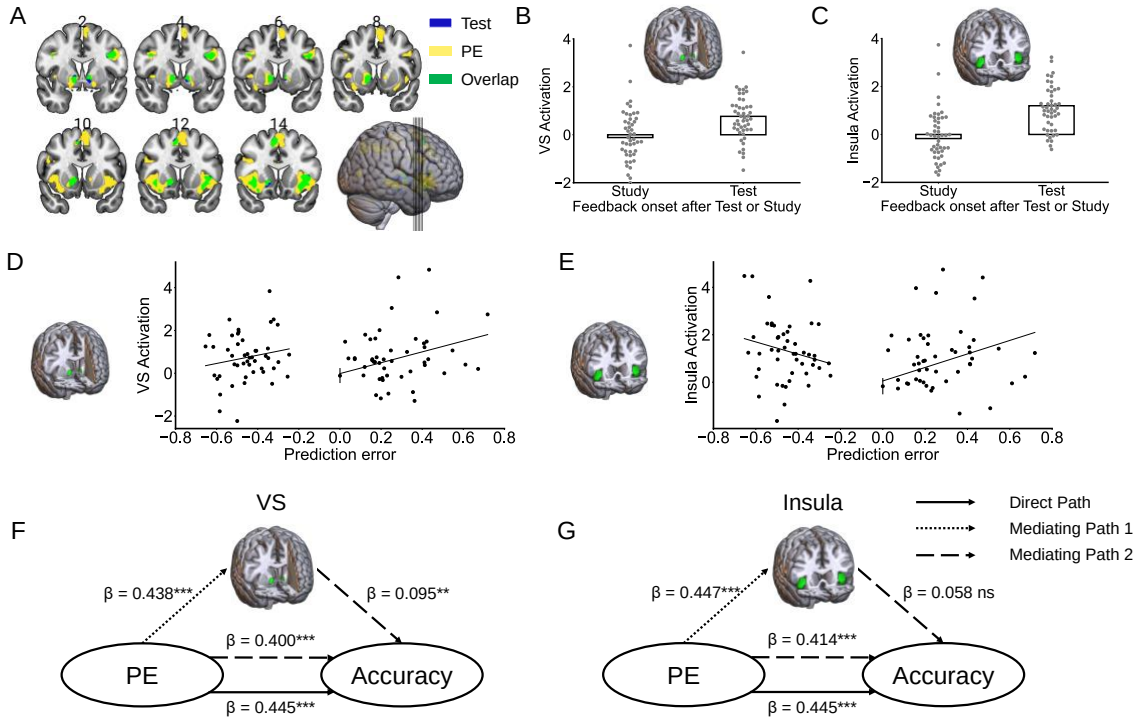

Figure 5: The overlapping neural activations of testing and prediction errors (A-E) and the mediating models for the prediction error effect (F-G). A: Testing and prediction errors elicit overlapping neural activations in the VS and insula. **B-C:** Overlapping VS and insula activation is stronger during feedback onsets after test than after study. **D:** The overlapping VS activations encode the original positive and negative prediction errors separately. **E:** The overlapping insula activations encode the absolute magnitudes of prediction errors. **F:** VS activation could mediate the prediction error effect on final assessment accuracy. **G:** Insula activation could not mediate the prediction error effect on final assessment accuracy. $p < 0.05$*, $p < 0.01$**, $p < 0.001$***, $p > 0.05$ ns.