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

Haopeng Chen (haopeng.chen@ugent.be)
Department of Experimental Psychology, Ghent University, 2 Henri Dunantlaan
Ghent, 9000 Belgium

Pieter Verbeke (pieter.verbeke@howest.be)
AI Lab, Howest University of Applied Sciences, 58 Marksesteenweg
Kortrijk, 8500 Belgium

Stefania Mattioni (stefania.mattioni@ugent.be)
Department of Experimental Psychology, Ghent University, 2 Henri Dunantlaan
Ghent, 9000 Belgium

Cristian Buc Calderon (cbuccald@gmail.com)
Centro Nacional de Inteligencia Artificial (CENIA), 4860 Av. Vicuña Mackenna
Santiago, 8320000, Chile

Tom Verguts (tom.verguts@ugent.be)
Department of Experimental Psychology, Ghent University, 2 Henri Dunantlaan
Ghent, 9000 Belgium

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) are more likely to be generated in testing (relative to in studying), testing can benefit more 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. Behaviorally and computationally, only a model incorporating predictive learning can account for the behavioral patterns and the robust testing effect. At the neural level, testing and prediction error both activate the canonical reward-related brain areas in the ventral striatum, insula, and midbrain. Crucially, back sorting analysis revealed that activation in the ventral striatum, insula, and midbrain can enhance declarative memory. These results provide strong and converging evidence for a predictive learning account of the testing effect.

Keywords: Testing effect; Predictive learning; fMRI; Ventral striatum; Insula.

Introduction

A remarkable finding is the testing effect, the robust phenomenon that testing enhances declarative memory retention more effectively than studying (Roediger & Karpicke, 2006). Accordingly, many learning apps such as Duolingo and Khanmigo are increasingly advocating users to have tests and retrieval, rather than restudy. However, the cognitive and neural origins of the testing effect remain a topic of debate. Here, we report and test an emergent predictive learning account (Chen, Hauspie, Ergo, Buc Calderon, & Verguts, 2025) as the cognitive and neural origin of the testing effect.

Predictive learning implies that minimizing prediction errors is a key objective for learning (Sutton & Barto, 2018). 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 (Calderon et al., 2021). Indeed, several studies have demonstrated that prediction errors can significantly promote declarative memory (Ergo, De Loof, & Verguts, 2020). 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 are more likely to appear in testing, not in studying (Chen et al., 2025). Indeed, testing provides more opportunities to predict possible answers, which are absent in mere studying. The mismatch between predictions and subsequent feedback generates prediction errors driving learning. Besides, prediction errors are proven to localize in the midbrain across animal species including rodents (Eshel, Tian, Bukwich, & Uchida, 2016), macaques (Schultz, Dayan, & Montague, 1997), and humans (Daniel & Pollmann, 2012), where they are encoded by dopamine 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 (VS) (Calderon et al., 2021), a key region for dopamine release. Based on these results, we propose a novel dopamine neural basis (VS) 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 and predictive learning share the same neural basis in the VS.

Methods

Experiment

Our declarative memory task consisted of four phases designed to help participants learn 90 Dutch-Swahili word pairs in the MRI scanner (Figure 1A). 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. Their accuracies and confidence ratings could be used to identify the learning status before formal manipulations. 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 feedback with the correct answer. Phase 4 consisted of two final assessments without feedback, with the same procedure as that of Phase 2.

Model simulations

We developed an associative memory neural network with an English input layer and a Swahili output layer, with each unit representing an English or Swahili word (Figure 1B). This model first initialize the connections between English and Swahili units by the equation below:

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

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.

After initialization, this model would receive testing or studying trials and implement predictive learning or/and Hebbian learning (a passive learning principle without prediction) to update the connections, followed by a final assessment.

On studying trials, the model could only implement Hebbian

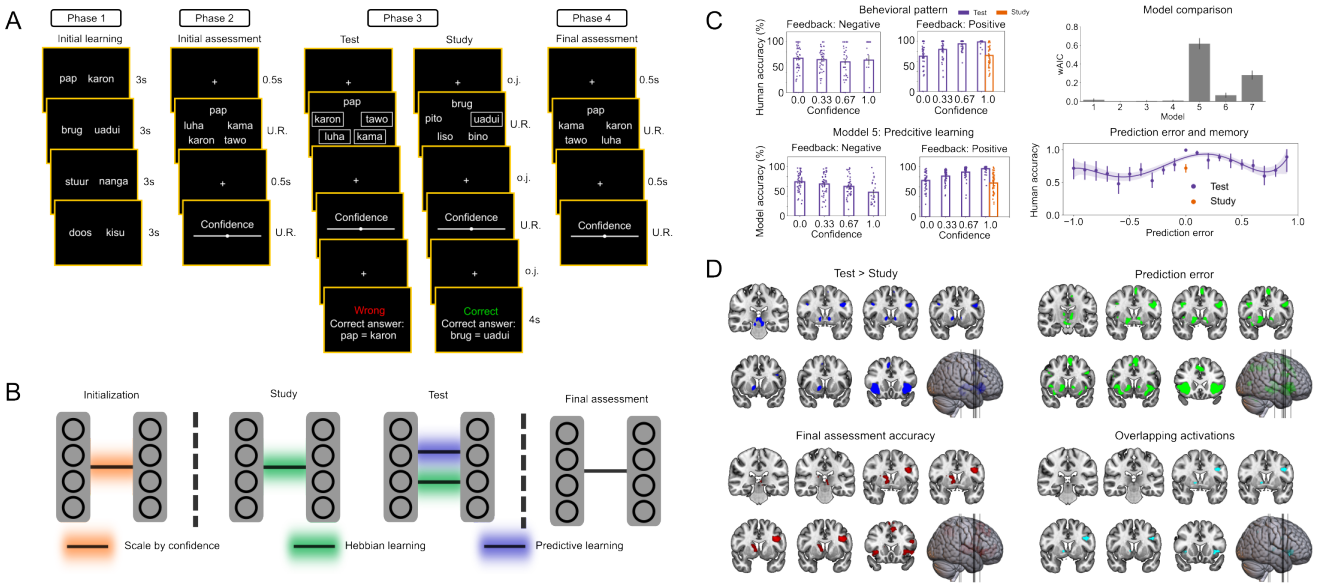

Figure 1: A, Experimental design. B, Model architecture (Model 1: Initial learning; Model 2: Hebbian learning; Model 3: Predictive learning; Model 4: Initial + Hebbian learning; Model 5: Initial + predictive learning; Model 6: Hebbian + predictive learning; Model 7: Full model). C, Behavioral results and model simulations. D, Test (vs. study), prediction error, and final assessment accuracy share the same neural basis in the ventral striatum and insula

learning by the equation below, as no predictions are involved in these trials.

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

In this equation, the model updated the weight w_{ij} only when a reward r_j was present (i.e., when feedback was positive). In contrast, testing trials could implement either Hebbian or predictive (or both) learning, with predictive learning implemented by the equation below:

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

Here, the model updated the weight w_{ij} based on the error between the actual feedback y_j and the model's prediction $\hat{y_j}$. A total of 7 models were built based on all possible combinations of initial learning, Hebbian learning, and predictive learning.

Results and Discussion

Behavioral patterns are depicted in Figure 1C (top-left panel), which replicate the robust testing effect ($\chi^2(1, N = 48) = 6.382, p = .012$). Besides, the testing effect became stronger ($\chi^2(1, N = 48) = 81.702, p < .001$) after controlling feedback and confidence ratings in Phase 3. Importantly, model comparison (Figure 1C; top-right panel) by wAIC (larger means better) suggests that the best fitting model is the model with initial learning and predictive learning (Model 5). Indeed, the model with initial learning and predictive learning successfully mimic the human behavioral pattern and the incorporated testing effect (Figure 1C; bottom-left panel). Additionally, we estimated the trial-level prediction errors (rounded to one decimal place) using Model 5. The estimated prediction errors influence the final assessment accuracy in a W-shaped pattern (Figure 1C; bottom-right panel), indicating that both highly positive and highly negative prediction errors can enhance declarative memory. The high final assessment accuracy in the zero prediction error condition may just reflect good initial learning. Finally, the fMRI analysis (Figure 1D) suggests that both testing (vs. studying) and prediction errors (estimated by Model 5) can trigger the VS, insula, and midbrain activations during feedback onset (Phase 3). Moreover, Stronger activations in the VS, insula, and midbrain during feedback onset are associated with correct, as opposed to incorrect, final assessments in Phase 4. Importantly, testing, prediction error, and final assessment related brain regions overlap in the VS and insula.

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. 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 (Bertsch, Pesta, Wiscott, & McDaniel, 2007), problem-based learning (Wood, 2003), and error-driven learning (Butterfield & Metcalfe, 2001).

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.

References

Bertsch, S., Pesta, B. J., Wiscott, R., & McDaniel, M. A. (2007). The generation effect: A meta-analytic review [Journal Article]. *Memory and Cognition*, *35*(2), 201-210. doi: 10.3758/BF03193441

Butterfield, B., & Metcalfe, J. (2001). Errors committed with high confidence are hypercorrected [Journal Article]. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *27*(6), 1491. doi: 10.1037//0278-7393.27.6.1491

Calderon, C. B., De Loof, E., Ergo, K., Snoeck, A., Boehler, C. N., & Verguts, T. (2021). Signed reward prediction errors in the ventral striatum drive episodic memory [Journal Article]. *J Neurosci*, *41*(8), 1716-1726. doi: 10.1523/JNEUROSCI.1785-20.2020

Chen, H., Hauspie, C., Ergo, K., Buc Calderon, C., & Verguts, T. (2025). Predictive learning as the basis of the testing effect [Journal Article]. *Communications Psychology*, *3*(1), 18. doi: 10.1038/s44271-025-00200-1

Daniel, R., & Pollmann, S. (2012). Striatal activations signal prediction errors on confidence in the absence of external feedback [Journal Article]. *NeuroImage*, *59*(4), 3457-3467. doi: https://doi.org/10.1016/j.neuroimage.2011.11.058

Ergo, K., De Loof, E., & Verguts, T. (2020). Reward prediction error and declarative memory [Journal Article]. *Trends in Cognitive Sciences*, *24*(5), 388-397. doi: 10.1016/j.tics.2020.02.009

Eshel, N., Tian, J., Bukwich, M., & Uchida, N. (2016). Dopamine neurons share common response function for reward prediction error [Journal Article]. *Nature Neuroscience*, *19*(3), 479-486. doi: 10.1038/nn.4239

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

Schultz, W., Dayan, P., & Montague, P. R. (1997). A neural substrate of prediction and reward [Journal Article]. *Science*, *275*(5306), 1593-1599. doi: doi:10.1126/science.275.5306.1593

Sutton, R., & Barto, A. (2018). *Reinforcement learning: an introduction* [Book]. Cambridge, MA: MIT.

Wood, D. F. (2003). Problem based learning [Journal Article]. *BMJ*, *326*(7384), 328-330. doi: 10.1136/bmj.326.7384.328