

Predictive Learning as the Basis of the Testing Effect

Haopeng Chen¹, Cathy Hauspie¹, Kate Ergo², Cristian Buc Calderon³, Tom Verguts¹

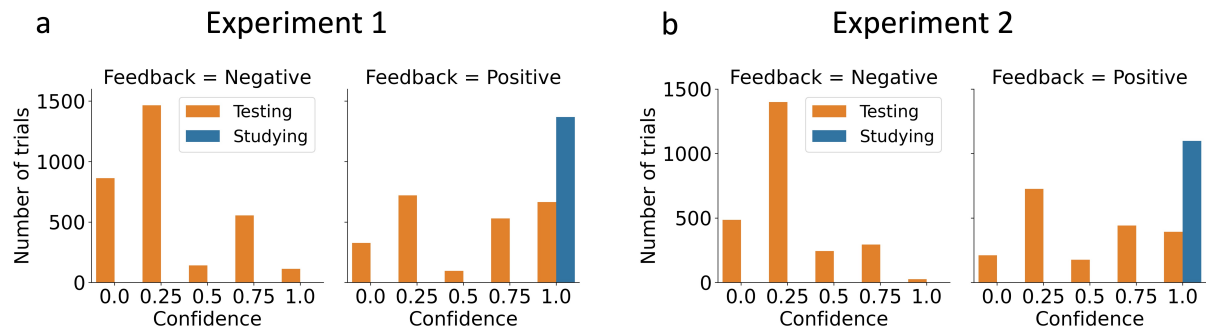
¹Department of Experimental Psychology, Faculty of Psychology and Educational Sciences,

Ghent University, Ghent, Belgium

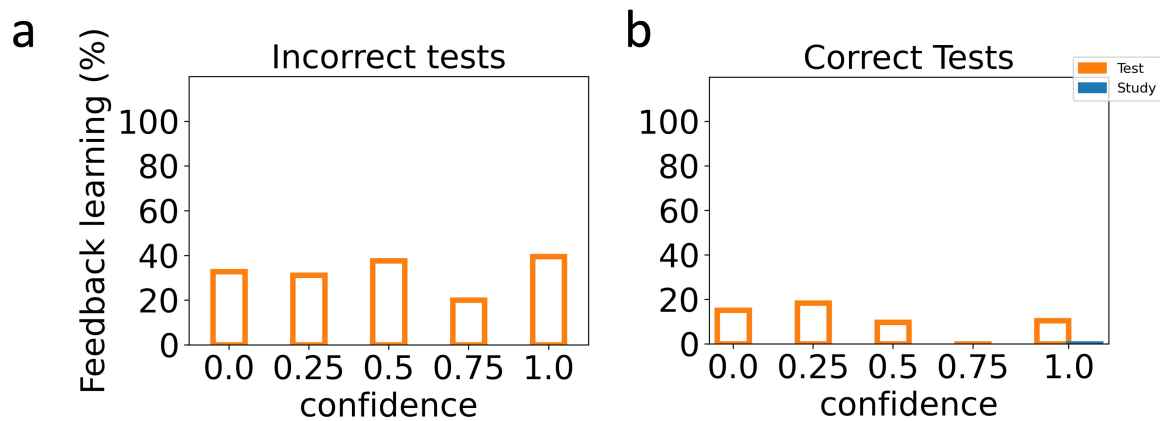
²Microsoft BeLux, Brussels, Belgium

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

Supplementary Information



Supplementary Figure 1 | Distribution of trials in each experimental condition. a. The distribution of trials in Experiment 1. **b.** The distribution of trials in Experiment 2.

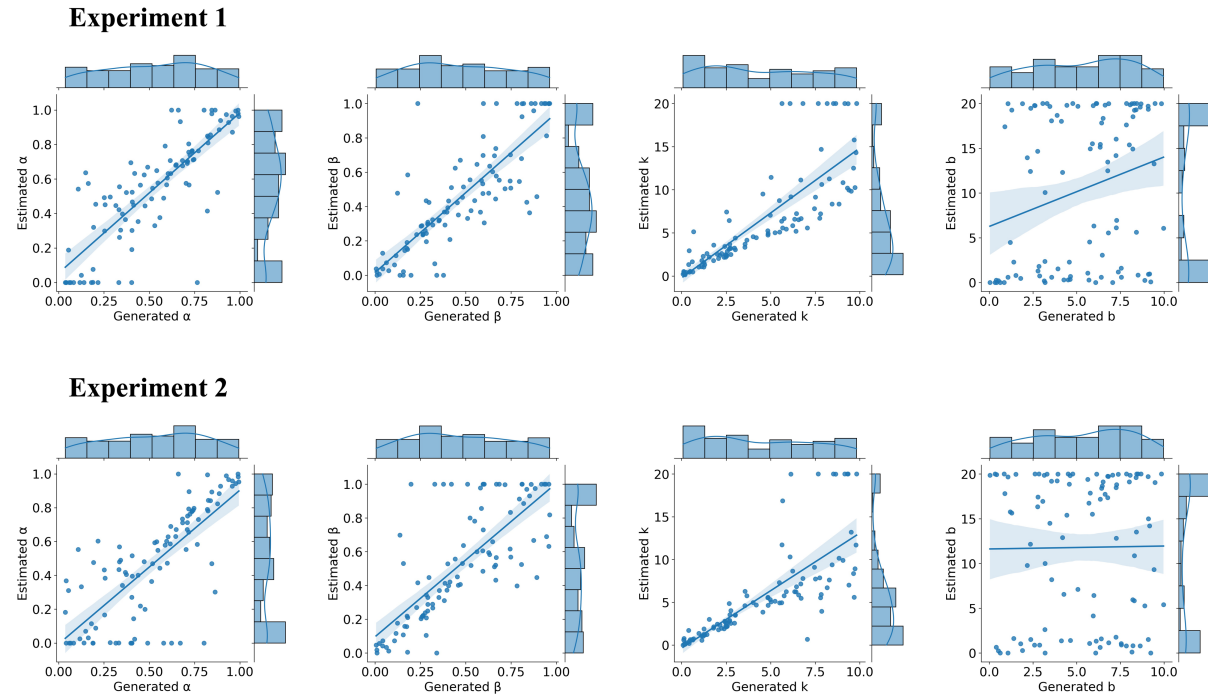


Supplementary Figure 2 | Pure learning from feedback. After subtracting the final recognition of data without feedback from those with feedback (mean value in each condition), the behavioural pattern revealed stronger learning from negative feedback compared to positive feedback. The x-axis depicts confidence ratings from Phase 3. **a:** Results for the incorrect tests in Phase 3; **b:** Results for the correct tests in Phase 3.

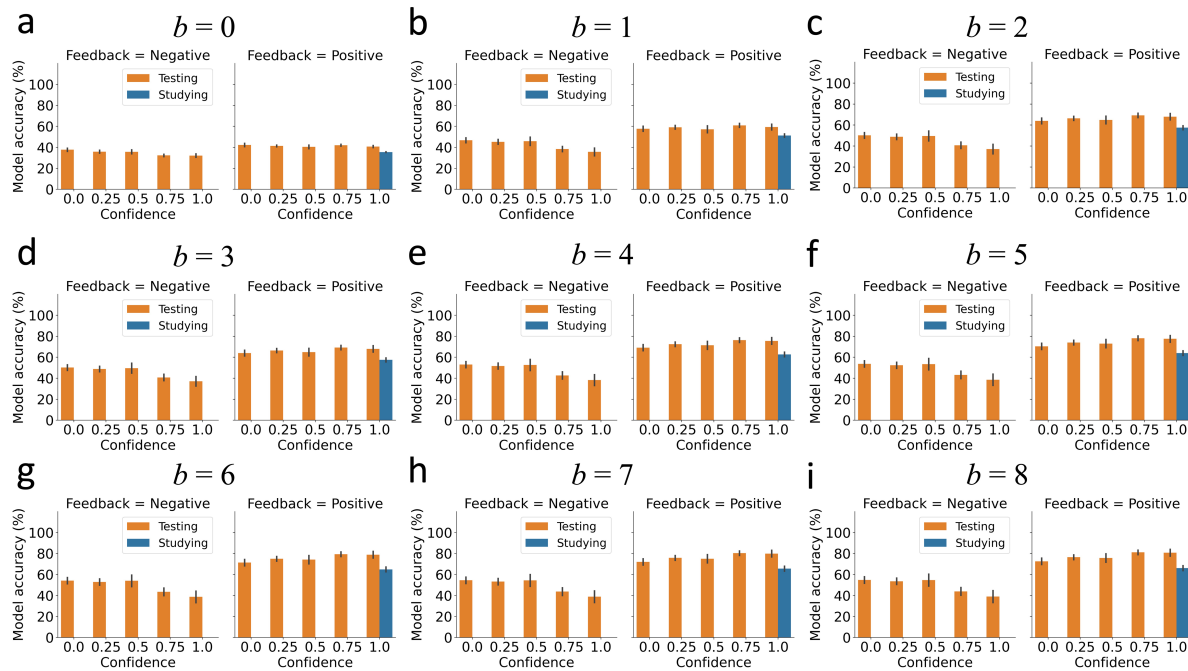
Supplementary Note 1: Parameter recovery

To check our model and algorithms, we carried out parameter recovery in three steps. First, a set of random parameters was generated from uniform distributions (α ranging uniformly from 0 to 1; $\beta = 1 - \alpha$; k and b ranging uniformly from 0 to 10). Second, using the generated parameters, a random participant's data was used to generate the choices in Phase 4 and Phase 5. Third, the model was used to fit the generated choices and yield estimated parameters. The above procedure was repeated 100 times. A strong correlation between generated and estimated parameters would indicate good parameter recovery¹.

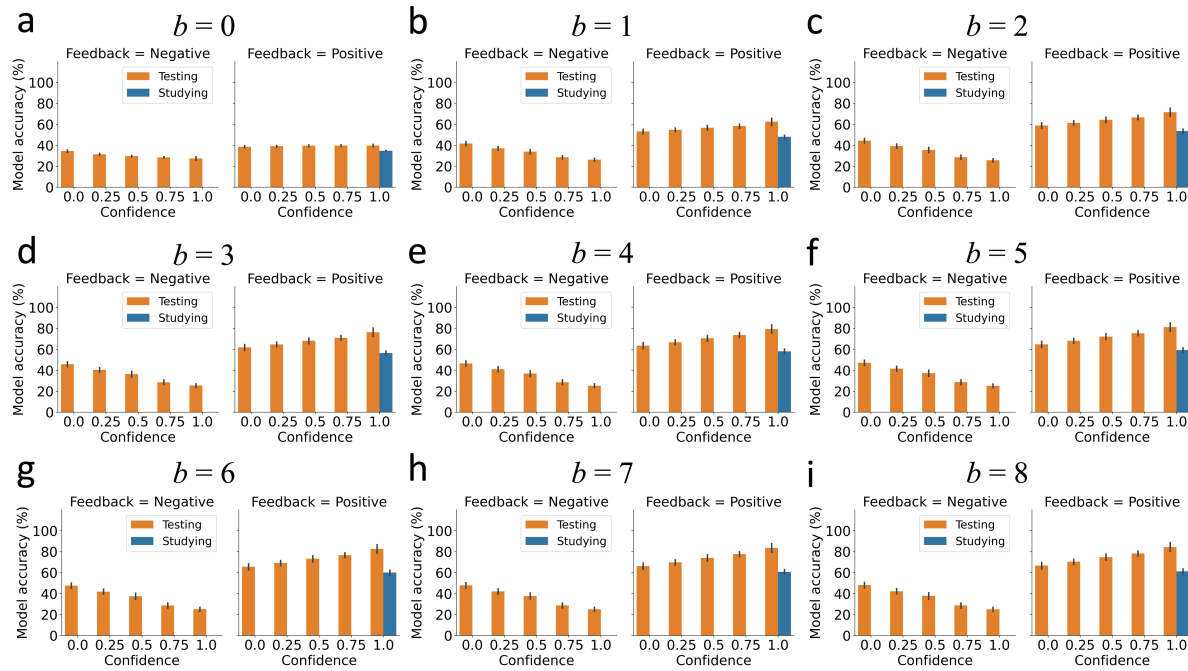
Because the full model (Model 7) is the most complex, we only report the parameter recovery of this model (the parameter recovery of other models was as good or better). As shown in Supplementary Figure 3, results from both Experiments showed that the recovery of α (Experiment1: $r = .820, p < .001, N = 100$ simulations; Experiment2: $r = .754, p < .001, N = 100$ simulations), β (Experiment1: $r = .820, p < .001, N = 100$ simulations; Experiment2: $r = .754, p < .001, N = 100$ simulations), and k (Experiment1: $r = .806, p < 0.001, N = 100$ simulations; Experiment2: $r = .772, p < 0.001, N = 100$ simulations) was very good, but the recovery of b was very bad (Experiment1: $r = .256, p = 0.010, N = 100$ simulations; Experiment2: $r = .011, p < .911, N = 100$ simulations). In further exploration, we found that the precise value of b is not important, because the model simulations did not change as long as b is higher than 3 (Supplementary Figures 4 and 5), which might explain the poor recovery of b . After all, the most important parameters are α and β , which are not impacted by the poor recovery of b .



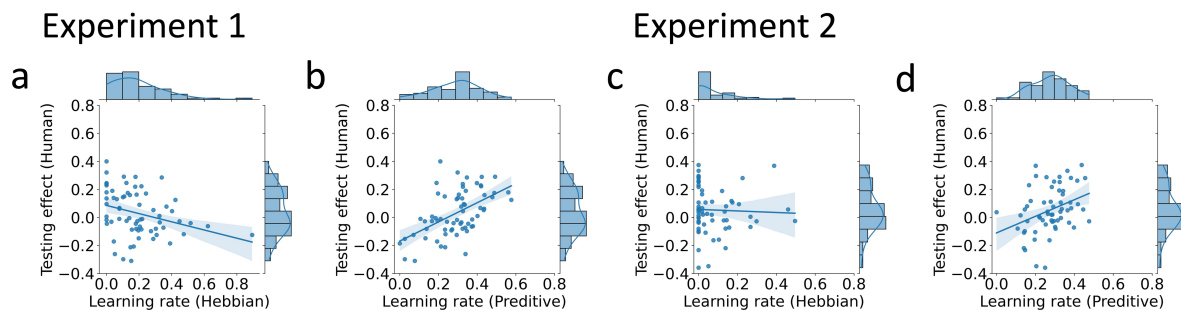
Supplementary Figure 3 | Parameter recovery of the most complex model (Model 7).
The y-axis represents the estimated parameters, the x-axis represents the true (randomly generated) parameters. Experiment 1: $N = 100$ simulations, Experiment 2: $N = 100$ simulations.



Supplementary Figure 4 | Simulating behavioural patterns by varying the parameter b in Experiment 1 ($N = 66$ simulations). The model simulations hardly change if b was higher than 3.



Supplementary Figure 5 | Simulating behavioural patterns by varying the parameter b in Experiment 2 ($N = 61$ simulations). The model simulations hardly change if b was higher than 3.

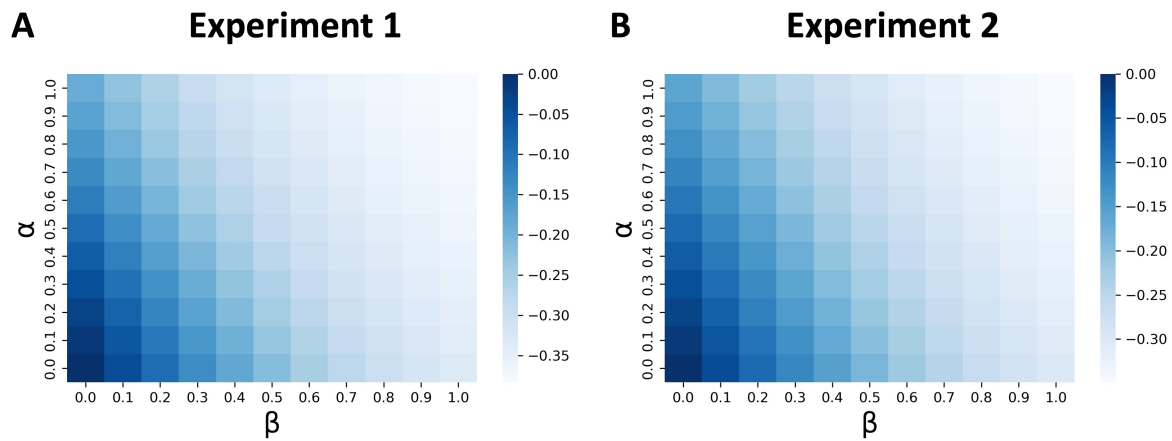


Supplementary Figure 6 | Predictive learning rate rather than Hebbian learning rate could predict the testing effect in human data (Full model with separate Hebbian and predictive learning rates). **a-b.** Predictive learning rate, rather than Hebbian learning rate could not predict the human testing effect in Experiment 1 ($\chi^2_h(1, N = 66) = 4.881, p = .027, \beta \pm 95\% CI = -.037 \pm .033, R^2 = .335; \chi^2_p(1, N = 66) = 22.696, p < .001, \beta \pm 95\% CI = .080 \pm .032, R^2 = .335$). **c-d.** Predictive learning rate, rather than Hebbian learning rate could predict the human testing effect in Experiment 2 ($\chi^2_h(1, N = 61) = .041, p = .839, \beta \pm 95\% CI = -.004 \pm .039, R^2 = .127; \chi^2_p(1, N = 61) = 8.317, p = .004, \beta \pm 95\% CI = .058 \pm .039, R^2 = .127$).

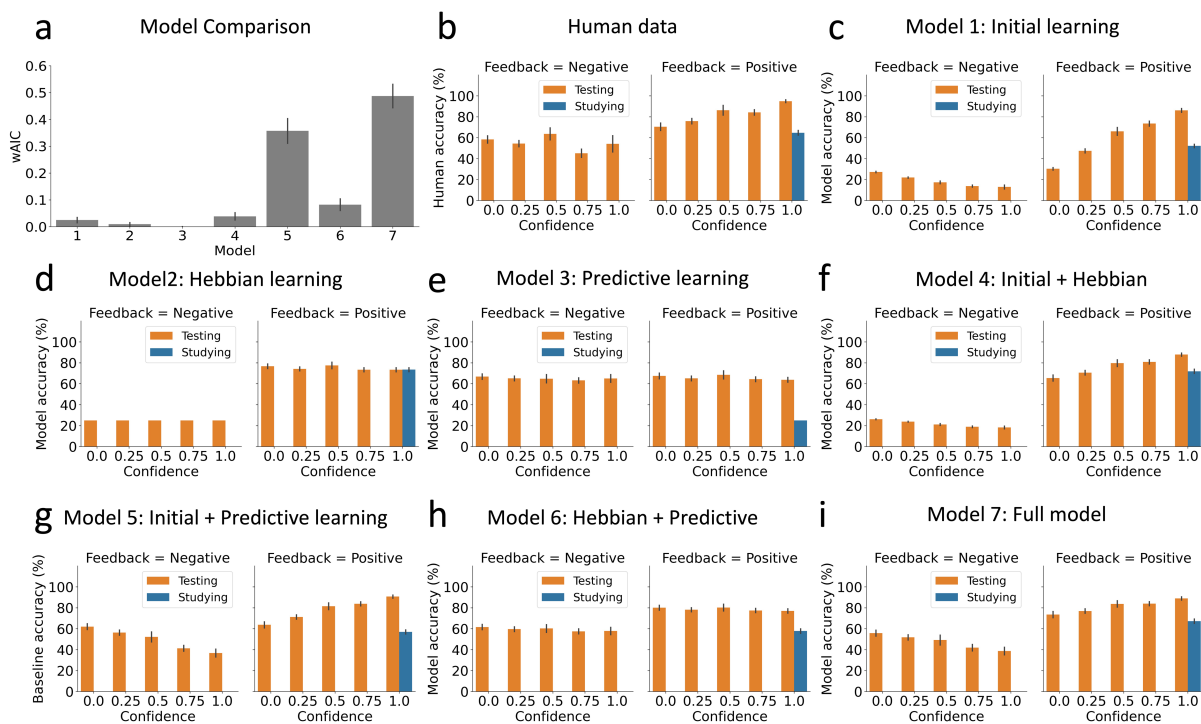
Supplementary Note 2: Can Hebbian learning produce a testing effect?

We tried to traverse the parameter spaces of parameters α and β ; parameters k and b were kept at their optimal (estimated) values. None of the parameter settings produced a

testing effect (Supplementary Figure 7). Indeed, the testing effects in the model were always negative, irrespective of which parameters were used.

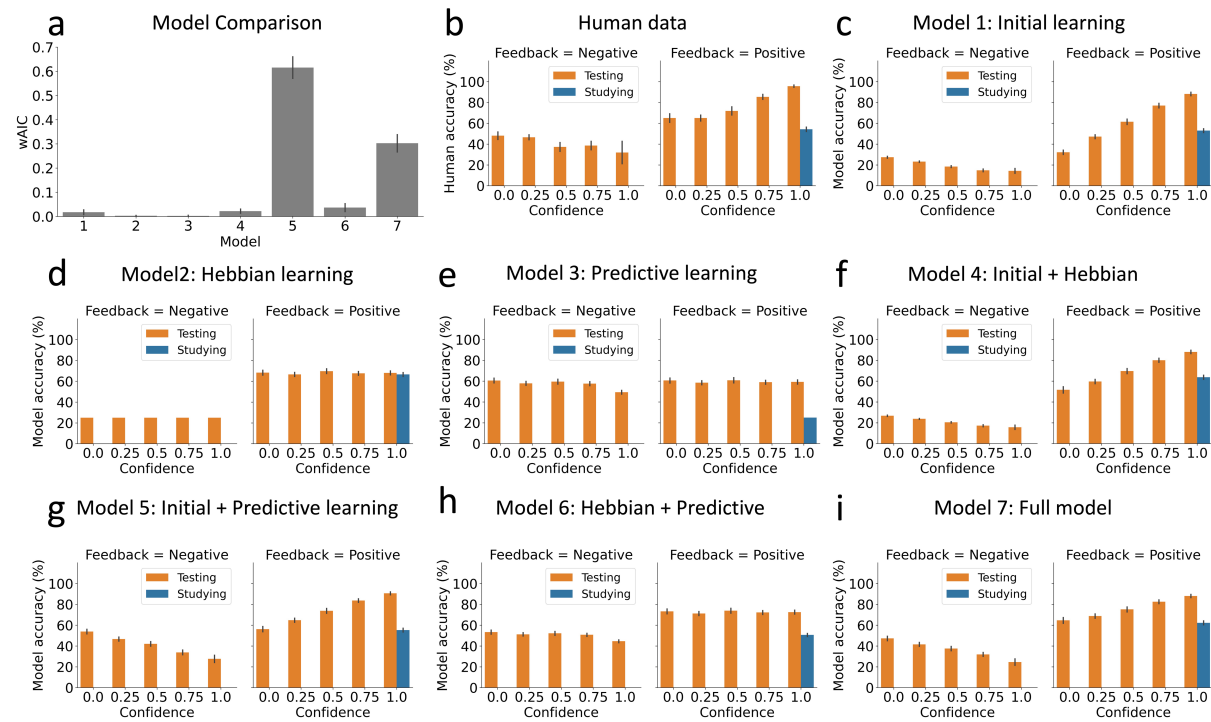


Supplementary Figure 7 | Model (4) simulated testing effect in parameter space. None of the parameter settings in Model 4 produced a testing effect. The color of the heat maps shows the magnitude of the testing effect (final accuracy after tests minus that after the study. e.g., a value of -0.10 means a reversed testing effect of 10%). The x-axis represents parameter β , and the y-axis represents parameter α . parameters k and b were kept at their optimal (estimated) values. Experiment 1: $N = 66$ simulations, Experiment 2: $N = 61$ simulations.

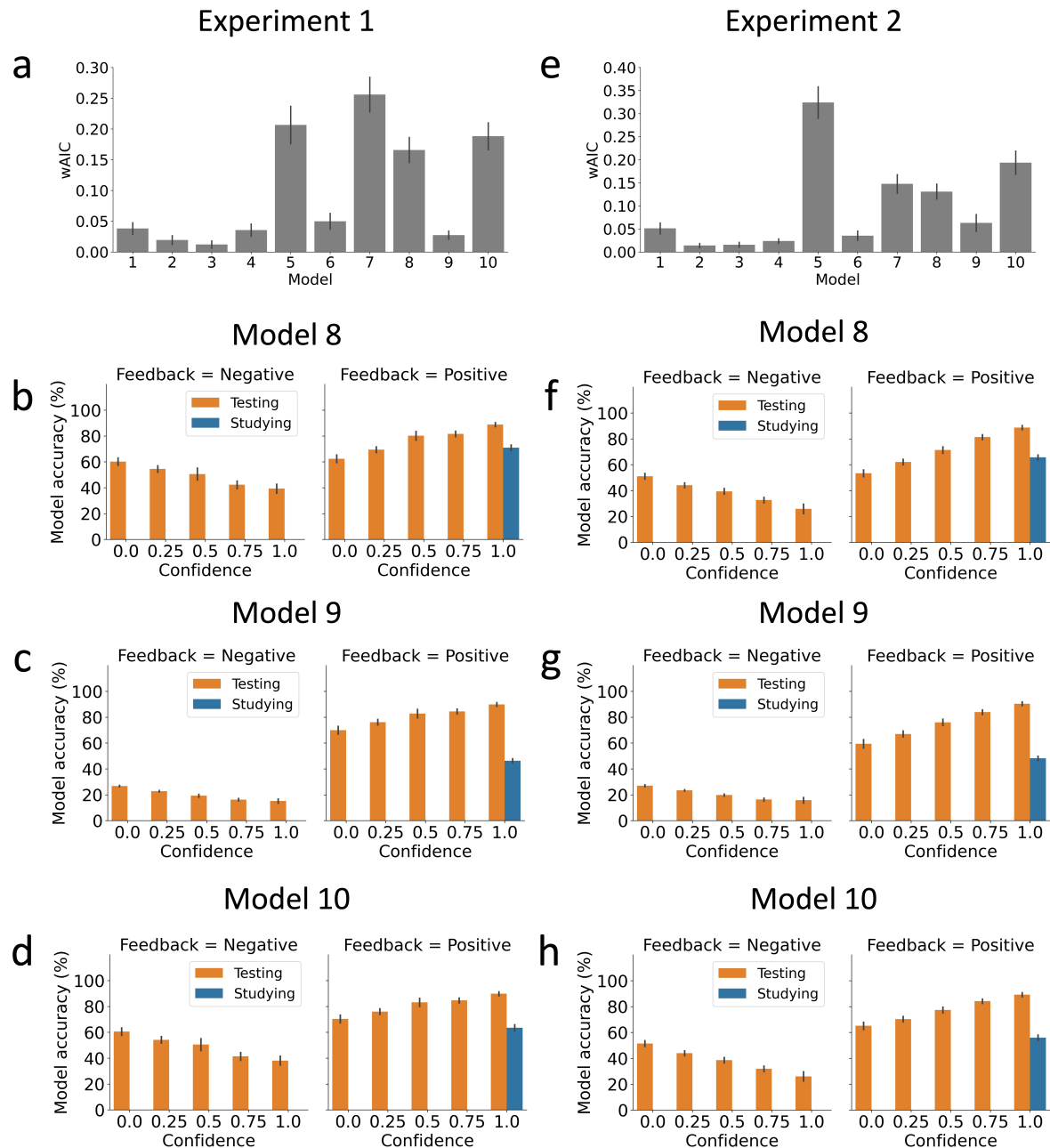


Supplementary Figure 8 | Model comparison and model simulations in Experiment 1 (all word pairs). **a.** Model comparison by wAIC. Higher wAIC means better model fitting. **b.** Human behavioral pattern. **c-i.** Simulations from all seven models suggested that models

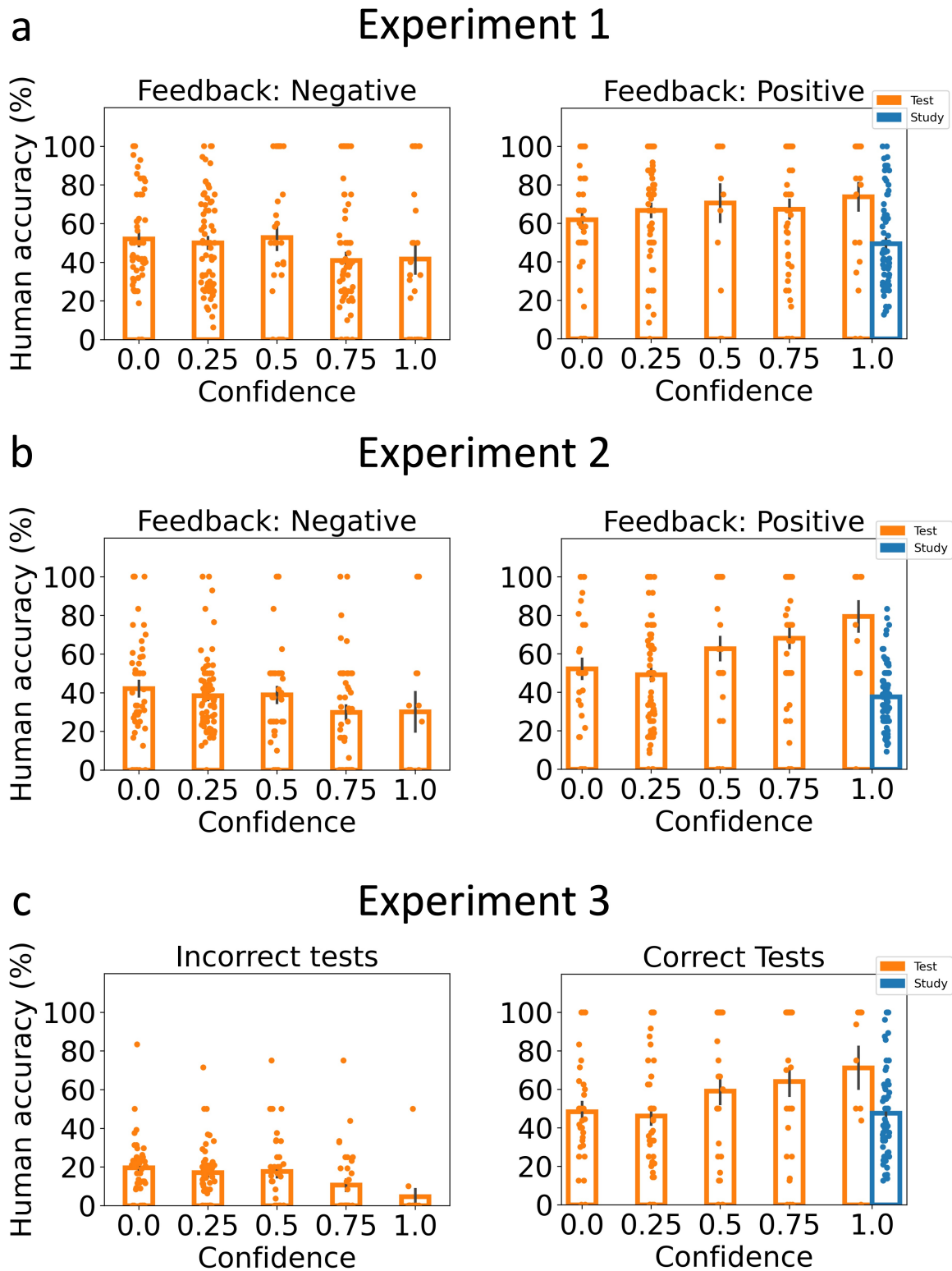
without predictive learning could not replicate the testing effect ($N = 66$ simulations). X-axis depicts confidence ratings from Phase 3; Error bars depict standard errors.



Supplementary Figure 9 | Model comparison and model simulations in Experiment 2 (all word pairs). **a.** Model comparison by wAIC. Higher wAIC means better model fitting. **b.** Human behavioral pattern. **c-i.** Simulations from all seven models suggested that models without predictive learning could not replicate the testing effect ($N = 61$ simulations). X-axis depicts confidence ratings from Phase 3; Error bars depict standard errors.

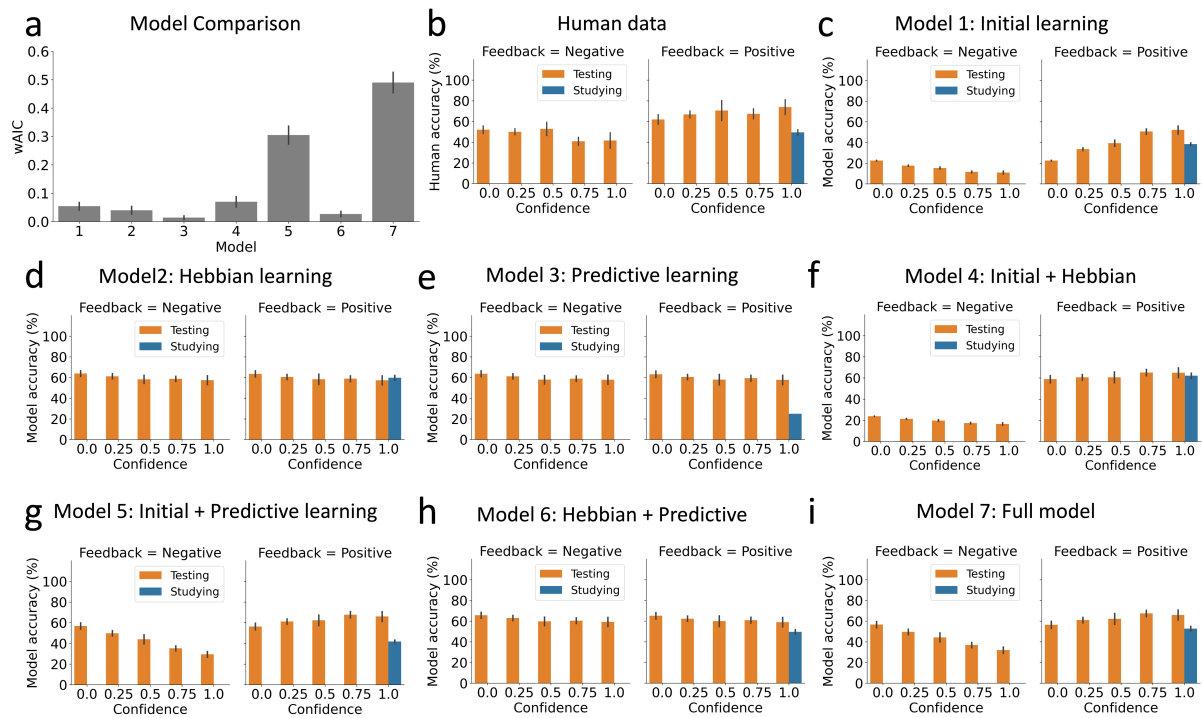


Supplementary Figure 10 | Further exploration of the Hebbian learning based testing effect (all word pairs). **a, e.** Model comparison in Experiment 1 and Experiment 2. Model 8: Initial learning and standard Hebbian learning without reward component. Model 9: Initial learning and Hebbian learning only with reward component in tests with positive feedback. Model 10: Initial learning; punishment-based Hebbian learning during tests with negative feedback, regular Hebbian learning during study, and reward-based Hebbian learning during tests with positive feedback. **b, f.** Simulations of model 8 in Experiment 1 and Experiment 2. **c, g.** Simulations of model 9 in Experiment 1 and Experiment 2. **d, h.** Simulations of model 10 in Experiment 1 and Experiment 2. X-axis depicts confidence ratings from Phase 3; Error bars depict standard errors. Experiment 1: N = 66 simulations, Experiment 2: N = 61 simulations.

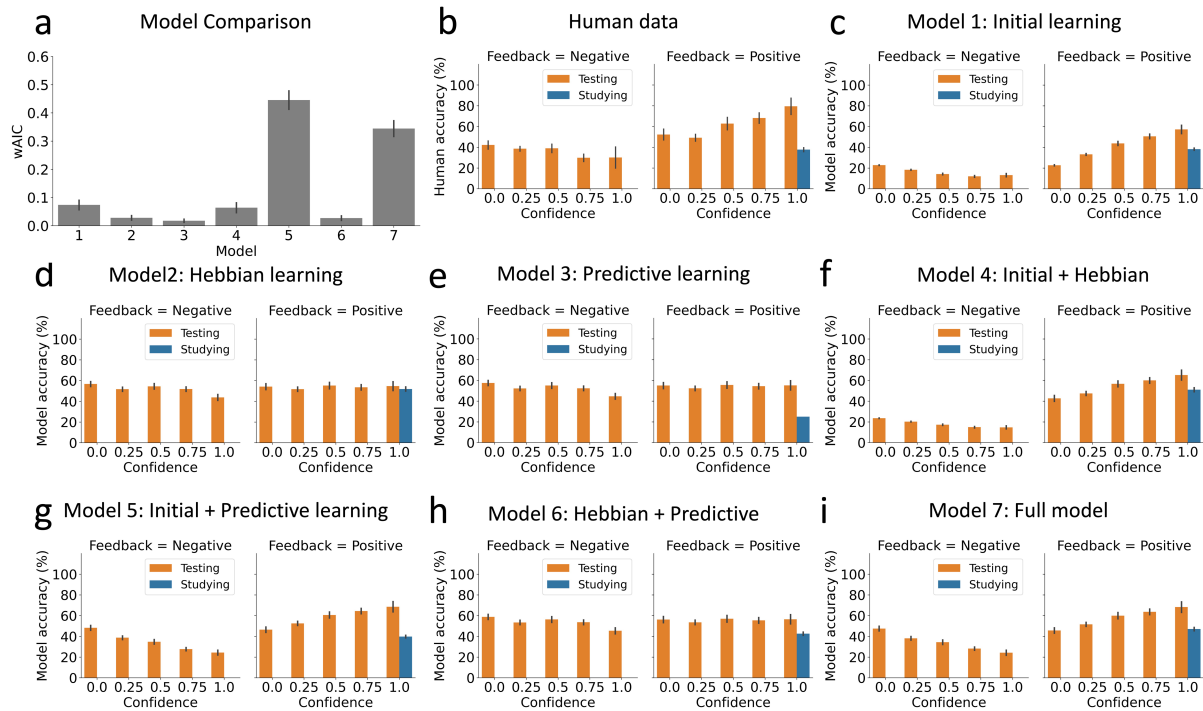


Supplementary Figure 11 | Behavioral patterns from three Experiments (all participants). **a.** Behavioral patterns from Experiment 1 replicated a significant testing effect in immediate recognitions; Final recognition after rewarded tests was better than after study, while unrewarded tests performed comparably to study (N = 76 participants). **b.** Behavioral patterns from Experiment 2 replicated the findings from Experiment 1 in a 24-hour delayed recognition phase (N = 76 participants). **c.** The testing effect was not significant in the data without feedback (immediate final recognition; N = 62 participants); The difference in final

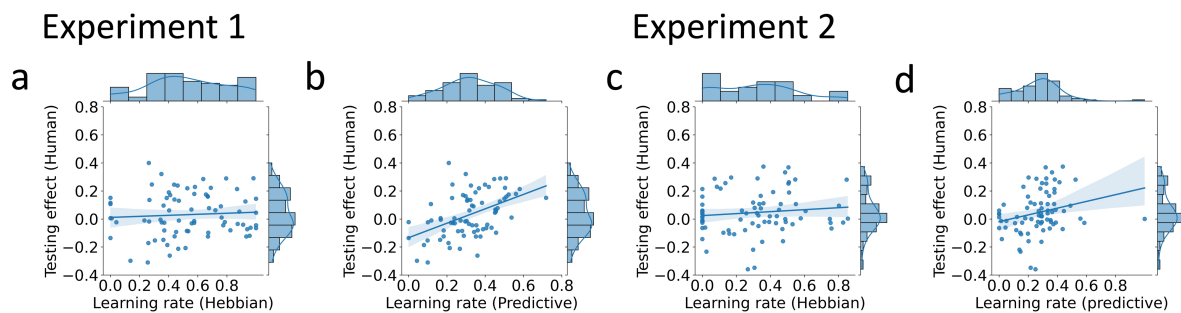
recognition after correct and incorrect tests was driven by initial learning rather than the
valence of feedback; Titles indicate correct and incorrect tests in Phase 3. X-axis depicts
confidence ratings from Phase 3; Error bars depict standard errors; Dots depict individual
data.



Supplementary Figure 12 | Model comparison and model simulations in Experiment 1 (all participants). **a.** Model comparison by wAIC. Higher wAIC means better model fitting. **b.** Human behavioral pattern. **c-i.** Simulations from all seven models suggested that models without predictive learning could not replicate the testing effect ($N = 76$ simulations). X-axis depicts confidence ratings from Phase 3; Error bars depict standard errors.



Supplementary Figure 13 | Model comparison and model simulations in Experiment 2 (all participants). **a.** Model comparison by wAIC. Higher wAIC means better model fitting. **b.** Human behavioral pattern. **c-i.** Simulations from all seven models suggested that models without predictive learning could not replicate the testing effect ($N = 76$ simulations). X-axis depicts confidence ratings from Phase 3; Error bars depict standard errors.



Supplementary Figure 14 | Predictive learning rate rather than Hebbian learning rate could predict the testing effect in human data (all participants). **a-b.** Predictive learning rate ($\chi^2(1, N = 76) = 21.628, p < .001, \beta \pm 95\% CI = .074 \pm .031, R^2 = .226$), rather than Hebbian learning rate ($\chi^2(1, N = 76) = .342, p = .558, \beta \pm 95\% CI = .011 \pm .035, R^2 = .005$) could predict the human testing effect in Experiment 1. **c-d.** Predictive learning rate ($\chi^2(1, N = 76) = 5.176, p = .023, \beta \pm 95\% CI = .039 \pm .034, R^2 = .065$), rather than Hebbian learning rate ($\chi^2(1, N = 76) = .956, p = .328, \beta \pm 95\% CI = .017 \pm .035, R^2 = .013$) could marginally predict the human testing effect in Experiment 2.

		N	Mean	SD	Median	Min	Max
Experiment 1	α	66	0.60	0.22	0.64	0.00	0.95
	β	66	0.40	0.22	0.36	0.05	1.00
	k	66	3.85	2.44	3.19	0.93	10.00
	b	66	6.53	4.37	9.79	0.05	10.00
Experiment 2	α	61	0.70	0.16	0.71	0.09	1.00
	β	61	0.30	0.16	0.29	0.00	0.91
	k	61	3.83	2.56	2.89	0.55	10.00
	b	61	6.65	4.40	9.73	0.00	10.00

Supplementary Table 1 | Parameter estimates of the best models in Experiment 1

(Model 7) and 2 (Model 5). N is the number of participants. Mean is the mean value across participants. SD is the standard deviation. Median is the median value across participants.

Min and max are minimum and maximum values, respectively.

184

Supplementary References

185 1. Wilson RC, Collins AG. Ten simple rules for the computational modeling of behavioral
186 data. *Elife* **8**, e49547 (2019).

187 <https://doi.org/10.7554/eLife.49547>

188

189