Supplementary Material for:

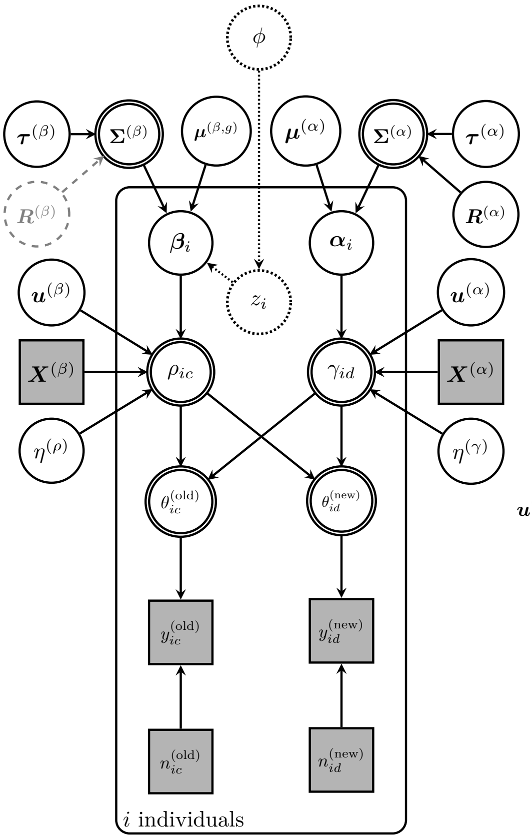
Blazej M. Baczkowski and Author2
Dep. Name, Uni Name

Blazej M. Baczkowski https://orcid.org/0000-0000-0000-0001
Author2 https://orcid.org/0000-0000-0000-0002
Correspondence concerning this article should be addressed to Blazej M. Baczkowski, Dep. Name, Uni Name, Address., City XX-XX, DE

Supplementary Figures

Figure S1

Graphical model representation of the Bayesian two-high-threshold model for old–new recognition judgments including latent mixture extension.


ϕ
$\boldsymbol{\tau}^{(\beta)}$
$\boldsymbol{\Sigma}^{(\beta)}$
$\boldsymbol{\mu}^{(\beta,g)}$
$\boldsymbol{\mu}^{(\alpha)}$
$\boldsymbol{\Sigma}^{(\alpha)}$
$\boldsymbol{\tau}^{(\alpha)}$
$\boldsymbol{R}^{(\beta)}$
$\boldsymbol{R}^{(\alpha)}$
$\boldsymbol{\beta}_i$
$\boldsymbol{\alpha}_i$
z_i
$\boldsymbol{u}^{(\beta)}$
$\boldsymbol{u}^{(\alpha)}$
$\boldsymbol{X}^{(\beta)}$
ρ_{ic}
γ_{id}
$\boldsymbol{X}^{(\alpha)}$
$\eta^{(\rho)}$
$\eta^{(\gamma)}$
$\theta_{ic}^{(\text{old})}$
$\theta_{id}^{(\text{new})}$
$y_{ic}^{(\text{old})}$
$y_{id}^{(\text{new})}$
$n_{ic}^{(\text{old})}$
$n_{id}^{(\text{new})}$
i individuals

$$
\begin{aligned}
\phi &\sim \text{Beta}(1,1) \quad \text{(mixture)} \\
z_i &\sim \text{Bernoulli}(\phi)^*,\ z_i \in \{0,1\} \quad \text{(mixture)} \\
\mu_1^{(\beta,g)} &\sim \begin{cases} \text{Normal}(-0.5,1.5),\ g=1 & \text{(baseline)} \\ \text{Normal}(-0.5,1.5),\ g \in \{0,1\} & \text{(mixture)} \end{cases} \\
\mu_j^{(\beta,g)} &\sim \begin{cases} \text{Normal}(0,1),\ j \in \{2{:}6\},\ g=1 \text{ (baseline)} \\ \text{Normal}(0,1),\ j \in \{2{:}6\},\ g \in \{0,1\} \text{ (mixture)} \end{cases} \\
\mu_1^{(\alpha)} &\sim \text{Normal}(-0.8,1.5) \\
\mu_2^{(\alpha)} &\sim \text{Normal}(0,1) \\
\tau_j^{(\beta)} &\sim \text{Gamma}(5,10),\ j \in \{1{:}6\} \\
\tau_k^{(\alpha)} &\sim \text{Gamma}(5,10),\ k \in \{1,2\} \\
\boldsymbol{R}^{(\beta)} &\sim \text{LKJcorr}(2),\ \boldsymbol{R}^{(\beta)} \in \mathbb{R}^{6\times 6} \\
\boldsymbol{R}^{(\alpha)} &\sim \text{LKJcorr}(2),\ \boldsymbol{R}^{(\alpha)} \in \mathbb{R}^{2\times 2} \\
\boldsymbol{\Sigma}^{(\beta)} &= \begin{cases} \text{diag}(\boldsymbol{\tau}^{(\beta)}) \cdot \boldsymbol{R}^{(\beta)} \cdot \text{diag}(\boldsymbol{\tau}^{(\beta)}) & \text{(baseline)} \\ \text{diag}(\boldsymbol{\tau}^{(\beta)})^2 & \text{(mixture)} \end{cases} \\
\boldsymbol{\Sigma}^{(\alpha)} &= \text{diag}(\boldsymbol{\tau}^{(\alpha)}) \cdot \boldsymbol{R}^{(\alpha)} \cdot \text{diag}(\boldsymbol{\tau}^{(\alpha)}) \\
\boldsymbol{\beta}_i \mid z_i &\sim \text{MvNormal}(\boldsymbol{\mu}^{(\beta,z_i)},\ \boldsymbol{\Sigma}^{(\beta)}),\ \boldsymbol{\beta}_i \in \mathbb{R}^{6\times 1} \\
\boldsymbol{\alpha}_i &\sim \text{MvNormal}(\boldsymbol{\mu}^{(\alpha)},\ \boldsymbol{\Sigma}^{(\alpha)}),\ \boldsymbol{\alpha}_i \in \mathbb{R}^{2\times 1} \\
\omega_j^{(\beta)},\ \omega_k^{(\alpha)} &\sim \text{Uniform}(0.01,0.4),\ j \in \{1{:}6\},\ k \in \{1,2\} \\
u_{j[study]}^{(\beta)} &\sim \text{Normal}(0,\omega_j^{(\beta)}),\ study \in \{1{:}4\} \\
u_{k[study]}^{(\alpha)} &\sim \text{Normal}(0,\omega_k^{(\alpha)}),\ study \in \{1{:}4\} \\
\boldsymbol{u}_{[study]}^{(\beta)},\ \boldsymbol{u}_{[study]}^{(\alpha)} &= [u_{1[study]}^{(\beta)},\dots,u_{6[study]}^{(\beta)}]^\top,\ [u_{1[study]}^{(\alpha)},u_{2[study]}^{(\alpha)}]^\top \\
c &= \begin{cases} 1,2,3 & \text{for phases 1–3}^{(\text{CS}-)} \\ 4,5,6 & \text{for phases 1–3}^{(\text{CS}+)} \end{cases} \\
d &= 1 \text{ for CS}-,\ 2 \text{ for CS}+ \\
w_c,\ w_d &= \begin{cases} -1/2 & \text{if } c,d \in \{\text{animals}\} \\ 1/2 & \text{if } c,d \in \{\text{tools}\} \end{cases} \\
\eta^{(\rho)},\ \eta^{(\gamma)} &\sim \text{Normal}(0,1) \\
\text{logit}(\rho_{ic}) &= \boldsymbol{X}_c^{(\beta)} \cdot (\boldsymbol{\beta}_i + \boldsymbol{u}_{[study]}^{(\beta)}) + w_c\ \eta^{(\rho)},\ \boldsymbol{X}_c^{(\beta)} \in \mathbb{R}^{1\times 6} \\
\text{logit}(\gamma_{id}) &= \boldsymbol{X}_d^{(\alpha)} \cdot (\boldsymbol{\alpha}_i + \boldsymbol{u}_{[study]}^{(\alpha)}) + w_d\ \eta^{(\gamma)},\ \boldsymbol{X}_d^{(\alpha)} \in \mathbb{R}^{1\times 2} \\
\theta_{ic}^{(\text{old})} &= \begin{cases} \rho_{ic} + (1-\rho_{ic})\ \gamma_{i1}, & c \in \{1,2,3\} \\ \rho_{ic} + (1-\rho_{ic})\ \gamma_{i2}, & c \in \{4,5,6\} \end{cases} \\
\theta_{id}^{(\text{new})} &= \begin{cases} (1-\rho_{i1}\ \rho_{i2}\ \rho_{i3})\ \gamma_{i1} & \text{if } d=1 \\ (1-\rho_{i4}\ \rho_{i5}\ \rho_{i6})\ \gamma_{i2} & \text{if } d=2 \end{cases} \\
y_{ic}^{(\text{old})} &\sim \text{Binomial}(\theta_{ic}^{(\text{old})},n_{ic}^{(\text{old})}),\ c \in \{1{:}6\} \\
y_{id}^{(\text{new})} &\sim \text{Binomial}(\theta_{id}^{(\text{new})},n_{id}^{(\text{new})}),\ d \in \{1,2\}
\end{aligned}
$$

Note. The left panel depicts a graphical model to illustrate the dependencies between observed data and latent parameters with nodes and directed edges. Nodes represent random variables, which can be either observed (shaded) or unobserved (unshaded), and either continuous (round) or discrete (square). Edges denote probabilistic or deterministic relationships. Solid nodes and edges are shared across both baseline and latent mixture models. The dashed grey node is present only in the baseline model, while dotted nodes appear exclusively in the mixture extension. Participant-level information is enclosed within a plate, indicating that the graphical structure is replicated across participants. The right panel details the generative model, specifying the assumed probability distributions over random variables and the deterministic equations that define their dependencies. The parameter ϕ represents the base rate of each qualitatively distinct latent participant group, and z indicates a discrete group assignment variable, shown here for clarity. In practice, z was marginalized out during model estimation.

Supplementary Methods

Latent-mixture two-high-threshold model (2HT): Exploring posterior group membership

Posterior group membership: presence across studies

To assess the generalizability and robustness of the two-group latent class solution across studies, we evaluated whether both groups were represented within each study. While the model included study-level random effects to capture between-study variation, we sought to confirm that all studies contributed participants to both latent subpopulations.

We extracted posterior group membership probabilities for each participant and summarized both hard class assignments and soft probabilities by study. This helped identify studies where one group may have been underrepresented or absent, potentially indicating sampling bias or limited within-study variability.

First, we calculated the proportion of participants assigned to each group (via hard assignments) per study to check for trivial or skewed distributions. Next, we plotted empirical cumulative distribution functions (ECDFs) of the average posterior probabilities to examine the separation between groups, expecting bimodal patterns near 0 and 1. Finally, we used the Kolmogorov–Smirnov (KS) statistic to compare the similarity of posterior probability distributions across studies. This statistic, defined as the maximum distance between two ECDFs, serves as a descriptive measure of the difference between distributions, ranging from 0 (identical) to 1 (completely distinct).

At each MCMC draw s we computed the KS distance between the posterior group membership probabilities for participants in two studies:

$$KS^{(s)} = D(p^{(s)}_{z[S_1]}, p^{(s)}_{z[S_2]})$$

where D denotes the distance between the two samples, and $p^{(s)}_{z[S]}$ is the vector of posterior probabilities of group membership for participants in particular study S at draw s. Given the sensitivity of the KS statistic to sample size, we consistently subsampled the larger study to match the sample size of the smaller study. This procedure yields a posterior distribution over KS statistics, allowing us to quantify and summarize uncertainty in the extent to which latent group membership distributions differ across studies.

Posterior group membership: association with participant-level covariates

We explored whether latent subgroup membership was associated with individual differences in learning rate as described by Rescorla-Wagner rule and anticipatory skin conductance response during conditioning.

Learning rate. We used trial-wise binary shock expectancy ratings during conditioning to estimate individual learning rates, based on a Rescorla-Wagner (RW) reinforcement learning model (Rescorla & Wagner, 1972; Tzovara et al., 2018):

$$x_t = x_{t-1} + \alpha(u_{t-1} - x_{t-1}).$$

Here, the associative strength x_t represents the predicted likelihood of an outcome and is updated based on the prediction error – the difference between the previous prediction x_{t-1} and the actual outcome u_{t-1}, where u_t is binary ($1 =$ shock, $0 =$ no shock). The prediction error is scaled by a subject-specific learning rate α ($0 < \alpha < 1$), which determines the rate of update based on outcomes of previous trials. In our setup, higher learning rates correspond to faster identification of the shock-predictive category, whereas lower learning rates correspond to greater uncertainty about the category-shock contingency.

We assumed that participants had no prior expectations about the CS+ and CS−, setting the initial associative strength to 0.5 for each condition, with a common learning rate. We interpreted the associative strength x_t as the participant's belief about the probability of the US occurring, which directly mapped onto the probability of a binary response indicating shock expectation, such that $\theta_t = x_t$. Hence, the behavioural binary response *shock* vs. *no-shock* was modelled as Bernoulli-distributed outcome:

$$y_t \sim \text{Bernoulli}(\theta_t).$$

We implemented the RW rule within a hierarchical framework, in which each participant had an individual learning rate α_i, modeled as a random variable drawn from a normal distribution with population expectation μ and variance τ^2 in the *logit* space:

$$\alpha_i \sim \text{Normal}(\mu, \tau).$$

We placed weakly informative priors on the population parameters:

$$\mu \sim \text{Normal}(0, 2),$$
$$\tau \sim \text{Gamma}(5, 10).$$

To obtain point estimates of the model parameters, we found the maximizing posterior density using the L-BFGS algorithm, implemented in the Stan language (Stan Development Team, 2023) and interfaced with the R package `rstan` (Stan Development Team, 2024).

To check for model mis-specification, we performed posterior predictive checks using graphical overlays and discrepancy measures based on simulated vs. observed summary statistics. Specifically, we computed the average proportion of shock expectancy across participants, aggregated within three trial bins per condition: trials 1–10, 11–20, and 21–30. Overall, the model tended to slightly underpredict the proportion of expected shocks compared to the observed data. However, these discrepancies were modest and fell within a reasonable range. The largest differences between predicted and observed means were approximately 0.67 vs. 0.72 for the CS+ condition and 0.03 vs. 0.12 for the CS– condition.

Visual inspection of trial-wise changes in associative strength (i.e., the model-derived probability of shock expectation) suggested that this underfitting may have stem from the use of a shared learning rate across conditions and the assumption of one-to-one correspondence between $\theta_t = x_t$. This constraint likely resulted in smoother predicted trajectories that failed to fully capture the more variable and less predictable patterns of behavioral responses observed in the empirical data.

Skin conductance response. The current study re-used condition-wise average anticipatory skin conductance responses (SCRs) during conditioning, as reported in the original study and provided by the corresponding author. Full details of the signal preprocessing and estimation procedures are described in the original report (Kalbe & Schwabe, 2021).

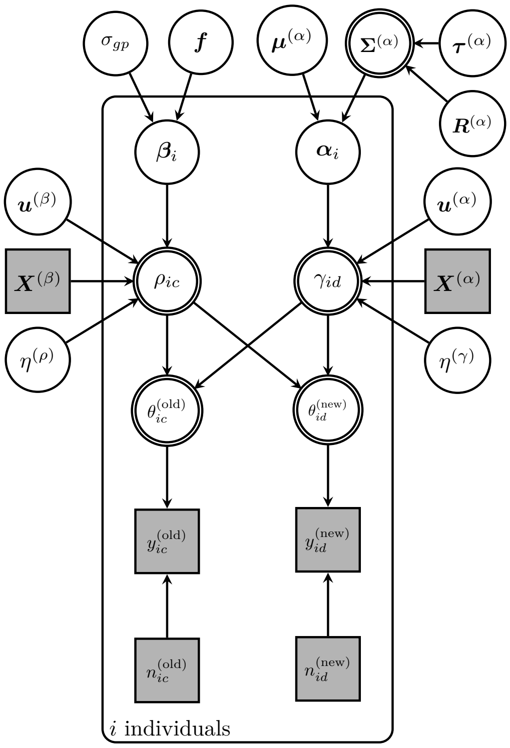
Analysis. To explore whether posterior group membership is associated with participant-level covariates, such as α learning rate of RW model or anticipatory SCR difference between CS+ vs. CS-, we performed a posterior-based association analysis inspired by the three-step approach of measuring the relationship between latent variables and covariates (Vermunt, 2010). After fitting the latent mixture model, we extracted the posterior probabilities of group membership for each participant $p_{z[i]}^{(s)}$, where s indexes posterior samples and i indexes participants. To assess the relationship between group membership and a covariate, we computed the Spearman rank-order correlation between the posterior probabilities of belonging to group 1 and the covariate values at each posterior draw:

$$r^{(s)} = \text{Spearman}(p_z^{(s)}, x)$$

where $p_z^{(s)} \in \mathbb{R}^n$ is the vector of probabilities indicating group 1 membership for all n participants at draw s and $x \in \mathbb{R}^n$ is the covariate vector. This yields a posterior distribution over the correlation coefficient r reflecting the uncertainty in both class membership and its association with the covariate.

Figure S2

Baseline two-high-threshold model: Gaussian Process extension.


σ_{gp}
$\boldsymbol{f}$
$\boldsymbol{\mu}^{(\alpha)}$
$\boldsymbol{\Sigma}^{(\alpha)}$
$\boldsymbol{\tau}^{(\alpha)}$
$\boldsymbol{R}^{(\alpha)}$
$\boldsymbol{\beta}_i$
$\boldsymbol{\alpha}_i$
$\boldsymbol{u}^{(\beta)}$
$\boldsymbol{u}^{(\alpha)}$
$\boldsymbol{X}^{(\beta)}$
ρ_{ic}
γ_{id}
$\boldsymbol{X}^{(\alpha)}$
$\eta^{(\rho)}$
$\eta^{(\gamma)}$
$\theta_{ic}^{(\text{old})}$
$\theta_{id}^{(\text{new})}$
$y_{ic}^{(\text{old})}$
$y_{id}^{(\text{new})}$
$n_{ic}^{(\text{old})}$
$n_{id}^{(\text{new})}$
i individuals

$$
\begin{array}{rcl}
\ell & \sim & \text{invGamma}(5,5) \\
\tau_{gp} & \sim & \text{invGamma}(5,5) \\
k(x|\tau_{gp},\ell) & = & \tau_{gp} \cdot \left(1 + \frac{\sqrt{3}\|x-x'\|}{\ell}\right) \exp\left(-\frac{\sqrt{3}\|x-x'\|}{\ell}\right) \\
f_j(x) & \sim & \text{MvNormal}(\mathbf{0}, k(x|\tau_{gp},\ell)),\ j \in \{1{:}6\} \\
\sigma_{gp} & \sim & \text{Gamma}(5,10) \\
\mu_1^{(\alpha)} & \sim & \text{Normal}(-0.8, 1.5) \\
\mu_2^{(\alpha)} & \sim & \text{Normal}(0,1) \\
\tau_j^{(\beta)} & \sim & \text{Gamma}(5,10),\ j \in \{1{:}6\} \\
\tau_k^{(\alpha)} & \sim & \text{Gamma}(5,10),\ k \in \{1,2\} \\
\boldsymbol{R}^{(\alpha)} & \sim & \text{LKJcorr}(2),\ \boldsymbol{R}^{(\alpha)} \in \mathbb{R}^{2\times 2} \\
\boldsymbol{\Sigma}^{(\alpha)} & = & \text{diag}(\boldsymbol{\tau}^{(\alpha)}) \cdot \boldsymbol{R}^{(\alpha)} \cdot \text{diag}(\boldsymbol{\tau}^{(\alpha)}) \\
\boldsymbol{\beta}_i & \sim & \text{MvNormal}(\boldsymbol{f}(x_i),\ \text{diag}(\sigma_{gp})^2),\ \boldsymbol{\beta}_i \in \mathbb{R}^{6\times 1} \\
\boldsymbol{\alpha}_i & \sim & \text{MvNormal}(\boldsymbol{\mu}^{(\alpha)},\ \boldsymbol{\Sigma}^{(\alpha)}),\ \boldsymbol{\alpha}_i \in \mathbb{R}^{2\times 1} \\
\omega_j^{(\beta)},\ \omega_k^{(\alpha)} & \sim & \text{Uniform}(0.01, 0.4),\ j \in \{1{:}6\},\ k \in \{1,2\} \\
u_{j[study]}^{(\beta)} & \sim & \text{Normal}(0, \omega_j^{(\beta)}),\ study \in \{1{:}4\} \\
u_{k[study]}^{(\alpha)} & \sim & \text{Normal}(0, \omega_k^{(\alpha)}),\ study \in \{1{:}4\} \\
\boldsymbol{u}_{[study]}^{(\beta)},\ \boldsymbol{u}_{[study]}^{(\alpha)} & = & [u_{1[study]}^{(\beta)}, \ldots, u_{6[study]}^{(\beta)}]^\top,\ [u_{1[study]}^{(\alpha)}, u_{2[study]}^{(\alpha)}]^\top \\
c & = & \begin{cases} 1,2,3 & \text{for phases } 1\text{–}3^{(\text{CS}-)} \\ 4,5,6 & \text{for phases } 1\text{–}3^{(\text{CS}+)} \end{cases} \\
d & = & 1 \text{ for CS}-,\ 2 \text{ for CS+} \\
w_c,\ w_d & = & \begin{cases} -1/2 & \text{if } c,d \in \{\text{animals}\} \\ 1/2 & \text{if } c,d \in \{\text{tools}\} \end{cases} \\
\eta^{(\rho)},\ \eta^{(\gamma)} & \sim & \text{Normal}(0,1) \\
\text{logit}(\rho_{ic}) & = & \boldsymbol{X}_c^{(\beta)} \cdot (\boldsymbol{\beta}_i + \boldsymbol{u}_{[study]}^{(\beta)}) + w_c\ \eta^{(\rho)},\ \boldsymbol{X}_c^{(\beta)} \in \mathbb{R}^{1\times 6} \\
\text{logit}(\gamma_{id}) & = & \boldsymbol{X}_d^{(\alpha)} \cdot (\boldsymbol{\alpha}_i + \boldsymbol{u}_{[study]}^{(\alpha)}) + w_d\ \eta^{(\gamma)},\ \boldsymbol{X}_d^{(\alpha)} \in \mathbb{R}^{1\times 2} \\
\theta_{ic}^{(\text{old})} & = & \begin{cases} \rho_{ic} + (1-\rho_{ic})\ \gamma_{i1}, & c \in \{1,2,3\} \\ \rho_{ic} + (1-\rho_{ic})\ \gamma_{i2}, & c \in \{4,5,6\} \end{cases} \\
\theta_{id}^{(\text{new})} & = & \begin{cases} (1-\rho_{i1}\ \rho_{i2}\ \rho_{i3})\ \gamma_{i1} & \text{if } d=1 \\ (1-\rho_{i4}\ \rho_{i5}\ \rho_{i6})\ \gamma_{i2} & \text{if } d=2 \end{cases} \\
y_{ic}^{(\text{old})} & \sim & \text{Binomial}(\theta_{ic}^{(\text{old})}, n_{ic}^{(\text{old})}),\ c \in \{1{:}6\} \\
y_{id}^{(\text{new})} & \sim & \text{Binomial}(\theta_{id}^{(\text{new})}, n_{id}^{(\text{new})}),\ d \in \{1,2\}
\end{array}
$$

Note. Graphical model of the extended 2HT model with a Gaussian Process (GP) prior over the memory recognition parameters, $\boldsymbol{\beta}_i$. Rather than sampling $\boldsymbol{\beta}_i$ directly from a multivariate normal distribution (as in the baseline model), this approach models them as noisy evaluations of a latent function $f(x)$ defined over subject-level covariates x (e.g., learning rate or anticipatory SCR). The GP captures smooth, nonlinear relationships between covariates and recognition performance. GP hyperparameters – kernel amplitude τ_{gp}, length scale ℓ, and observation noise σ_{gp} – are assigned weakly informative priors.

Baseline two-high-threshold model (2HT): Gaussian Process extension

To explore whether individual differences in memory recognition vary systematically with subject-level features – assuming a single generative process shared across all participants – we extended the baseline 2HT model. Specifically, we placed a Gaussian Process (GP) prior over the individual $\boldsymbol{\beta}_i \in \mathbb{R}^{6\times 1}$ parameters, allowing them to flexibly vary as a function of covariates such as the RW learning rate and anticipatory physiological arousal (see graphical model in Figure S2).

In the baseline model, the subject-specific parameters $\boldsymbol{\beta}_i$ are drawn from a multivariate normal distribution with parameter-specific population means:

$$\boldsymbol{\beta}_i \sim \text{MvNormal}(\boldsymbol{\mu}^{(\beta)}, \boldsymbol{\Sigma}^{(\beta)}).$$

In the GP-extended model, we instead model these weights as noisy evaluations of latent functions defined over a covariate space:

$$\boldsymbol{\beta}_i \sim \text{MvNormal}(\boldsymbol{f}(x_i), \text{diag}(\sigma_{gp}^2)),$$

where $\boldsymbol{f}(x_i) \in \mathbb{R}^6$ denotes the vector of function outputs evaluated at the covariate vector x_i for subject i, and σ_{gp}^2 captures independent Gaussian observation noise for each parameter.

Each of the six functions in $\boldsymbol{f}$ is modeled as a draw from a zero-mean GP with a shared Matérn 3/2 kernel:

$$f_j(x) \sim \text{GP}(0,\, k(x, x' \mid \tau_{gp}, \ell)), \quad \text{for } j = 1, \ldots, 6,$$
$$k(x, x' \mid \tau_{gp}, \ell) = \tau_{gp} \cdot \left(1 + \frac{\sqrt{3}\|x - x'\|}{\ell}\right) \exp\left(-\frac{\sqrt{3}\|x - x'\|}{\ell}\right),$$

where τ_{gp} controls the output scale and ℓ the length scale. This kernel captures moderate smoothness and is computationally efficient, making it well suited for small- to medium-scale modeling.

The input features x_i varied across two model variants. In one, $x_i \in \mathbb{R}$ represented the subject's learning rate from the RW model. In the other, $x_i \in \mathbb{R}^2$ consisted of anticipatory SCR values derived from two distinct estimation methods.

We placed standard weakly informative priors on the GP hyperparameters:

$$\begin{aligned} \ell &\sim \text{InvGamma}(5, 5) && \text{(length scale)} \\ \tau_{gp} &\sim \text{InvGamma}(5, 5) && \text{(kernel amplitude)} \\ \sigma_{gp} &\sim \text{Gamma}(5, 10) && \text{(observation noise).} \end{aligned}$$

This GP-based framework allowed us to assess whether memory recognition varies systematically across a latent space defined by either individual learning dynamics or anticipatory physiological arousal.

Bayesian one-high-threshold (1HT) model of memory recognition

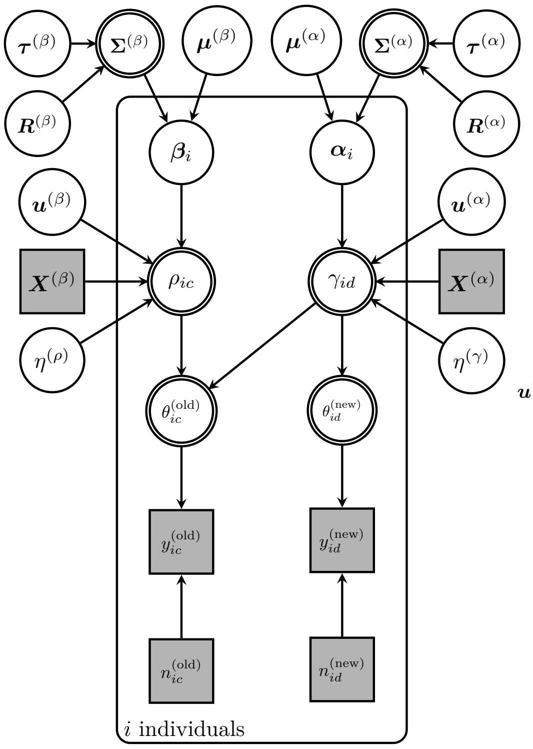
The one-high-threshold (1HT) recognition model shares the same underlying assumptions and decision-tree framework as the two-high-threshold (2HT) model, with a key distinction in the treatment of false alarms.

In the 2HT model, false alarms are attributed to a combination of detecting new (i.e., memory recognition) and guessing. In contrast, the 1HT model assumes that false alarms result exclusively from guessing, with no contribution from recognition processes. Hit rates remain a mixture of recognition and guessing. This implies that, under the current design, responses to new items are not affected by memory of previously encoded items from the same semantic category.

As a result, the 1HT model adopts the same data-generating assumptions and prior distribution specifications as the 2HT model (see Figure S3 for a graphical representation), with the sole

Figure S3

Graphical model representation of the Bayesian one-high-threshold model for old–new recognition judgments.


$\boldsymbol{\tau}^{(\beta)}$
$\boldsymbol{\Sigma}^{(\beta)}$
$\boldsymbol{\mu}^{(\beta)}$
$\boldsymbol{\mu}^{(\alpha)}$
$\boldsymbol{\Sigma}^{(\alpha)}$
$\boldsymbol{\tau}^{(\alpha)}$
$\boldsymbol{R}^{(\beta)}$
$\boldsymbol{R}^{(\alpha)}$
$\boldsymbol{\beta}_i$
$\boldsymbol{\alpha}_i$
$\boldsymbol{u}^{(\beta)}$
$\boldsymbol{u}^{(\alpha)}$
$\boldsymbol{X}^{(\beta)}$
ρ_{ic}
γ_{id}
$\boldsymbol{X}^{(\alpha)}$
$\eta^{(\rho)}$
$\eta^{(\gamma)}$
$\theta_{ic}^{(\text{old})}$
$\theta_{id}^{(\text{new})}$
$y_{ic}^{(\text{old})}$
$y_{id}^{(\text{new})}$
$n_{ic}^{(\text{old})}$
$n_{id}^{(\text{new})}$
i individuals

$$
\begin{aligned}
\mu_1^{(\beta)} &\sim \text{Normal}(-0.5, 1.5)\\
\mu_j^{(\beta)} &\sim \text{Normal}(0, 1),\ j \in \{2{:}6\}\\
\mu_1^{(\alpha)} &\sim \text{Normal}(-0.8, 1.5)\\
\mu_2^{(\alpha)} &\sim \text{Normal}(0, 1)\\
\tau_j^{(\beta)} &\sim \text{Gamma}(5, 10),\ j \in \{1{:}6\}\\
\tau_k^{(\alpha)} &\sim \text{Gamma}(5, 10),\ k \in \{1, 2\}\\
\boldsymbol{R}^{(\beta)} &\sim \text{LKJcorr}(2),\ \boldsymbol{R}^{(\beta)} \in \mathbb{R}^{6\times 6}\\
\boldsymbol{R}^{(\alpha)} &\sim \text{LKJcorr}(2),\ \boldsymbol{R}^{(\alpha)} \in \mathbb{R}^{2\times 2}\\[1ex]
\boldsymbol{\Sigma}^{(\beta)} &= \text{diag}(\boldsymbol{\tau}^{(\beta)}) \cdot \boldsymbol{R}^{(\beta)} \cdot \text{diag}(\boldsymbol{\tau}^{(\beta)})\\
\boldsymbol{\Sigma}^{(\alpha)} &= \text{diag}(\boldsymbol{\tau}^{(\alpha)}) \cdot \boldsymbol{R}^{(\alpha)} \cdot \text{diag}(\boldsymbol{\tau}^{(\alpha)})\\[1ex]
\boldsymbol{\beta}_i &\sim \text{MvNormal}(\boldsymbol{\mu}^{(\beta)},\ \boldsymbol{\Sigma}^{(\beta)}),\ \boldsymbol{\beta}_i \in \mathbb{R}^{6\times 1}\\
\boldsymbol{\alpha}_i &\sim \text{MvNormal}(\boldsymbol{\mu}^{(\alpha)},\ \boldsymbol{\Sigma}^{(\alpha)}),\ \boldsymbol{\alpha}_i \in \mathbb{R}^{2\times 1}\\[1ex]
\omega_j^{(\beta)},\ \omega_k^{(\alpha)} &\sim \text{Uniform}(0.01, 0.4),\ j \in \{1{:}6\},\ k \in \{1, 2\}\\
u_{j[study]}^{(\beta)} &\sim \text{Normal}(0, \omega_j^{(\beta)}),\ study \in \{1{:}4\}\\
u_{k[study]}^{(\alpha)} &\sim \text{Normal}(0, \omega_k^{(\alpha)}),\ study \in \{1{:}4\}\\
\boldsymbol{u}_{[study]}^{(\beta)},\ \boldsymbol{u}_{[study]}^{(\alpha)} &= [u_{1[study]}^{(\beta)}, \ldots, u_{6[study]}^{(\beta)}]^\top,\ [u_{1[study]}^{(\alpha)}, u_{2[study]}^{(\alpha)}]^\top\\[1ex]
c &= \begin{cases} 1, 2, 3 & \text{for phases 1–3}^{(\text{CS}-)}\\ 4, 5, 6 & \text{for phases 1–3}^{(\text{CS}+)} \end{cases}\\
d &= 1 \text{ for CS}-,\ 2 \text{ for CS}+\\
w_c,\ w_d &= \begin{cases} -1/2 & \text{if } c, d \in \{\text{animals}\}\\ 1/2 & \text{if } c, d \in \{\text{tools}\} \end{cases}\\[1ex]
\eta^{(\rho)},\ \eta^{(\gamma)} &\sim \text{Normal}(0, 1)\\[1ex]
\text{logit}(\rho_{ic}) &= \boldsymbol{X}_c^{(\beta)} \cdot (\boldsymbol{\beta}_i + \boldsymbol{u}_{[study]}^{(\beta)}) + w_c\ \eta^{(\rho)},\ \boldsymbol{X}_c^{(\beta)} \in \mathbb{R}^{1\times 6}\\
\text{logit}(\gamma_{id}) &= \boldsymbol{X}_d^{(\alpha)} \cdot (\boldsymbol{\alpha}_i + \boldsymbol{u}_{[study]}^{(\alpha)}) + w_d\ \eta^{(\gamma)},\ \boldsymbol{X}_d^{(\alpha)} \in \mathbb{R}^{1\times 2}\\[1ex]
\theta_{ic}^{(\text{old})} &= \begin{cases} \rho_{ic} + (1 - \rho_{ic})\ \gamma_{i1}, & c \in \{1, 2, 3\}\\ \rho_{ic} + (1 - \rho_{ic})\ \gamma_{i2}, & c \in \{4, 5, 6\} \end{cases}\\
\theta_{id}^{(\text{new})} &= \gamma_{id}\\[2ex]
y_{ic}^{(\text{old})} &\sim \text{Binomial}(\theta_{ic}^{(\text{old})}, n_{ic}^{(\text{old})}),\ c \in \{1{:}6\}\\
y_{id}^{(\text{new})} &\sim \text{Binomial}(\theta_{id}^{(\text{new})}, n_{id}^{(\text{new})}),\ d \in \{1, 2\}
\end{aligned}
$$

Note. Unlike the 2HT model, where false alarms arise from memory recognition and guessing, this model assumes false alarms arise solely from guessing, excluding any recognition influence.

difference being in the modeling of false alarm rates. Specifically, false alarms are modeled using a binomial likelihood function, where the parameter $\theta_{id}^{(\text{new})}$ is set equal to the guessing parameter γ_{id}:

$$\theta_{id}^{(\text{new})} = \gamma_{id},$$

where d indicates semantic category assigned to either CS- or CS+ condition.

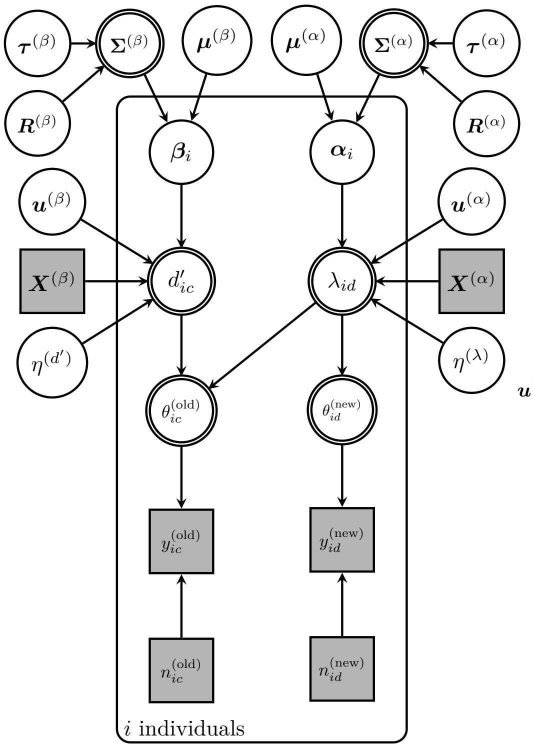
To evaluate model adequacy and potential mis-specification, we conducted posterior predictive checks. These included comparisons between observed and simulated data for summary statistics – specifically, grand means of hits and false alarms – aggregated both across and within studies.

Bayesian model of memory recognition based on signal detection theory (SDT)

The signal detection theory model of memory recognition (SDT) belongs to a different family than threshold models, treating memory strength as a continuous variable. While threshold models assume categorical, all-or-none representations, SDT treats memory strength as continuous, with recognition occurring when strength exceeds a decision criterion.

Figure S4

Graphical model representation of the Bayesian model for old–new recognition judgments based on signal detection theory.


$\boldsymbol{\tau}^{(\beta)}$
$\boldsymbol{\Sigma}^{(\beta)}$
$\boldsymbol{\mu}^{(\beta)}$
$\boldsymbol{\mu}^{(\alpha)}$
$\boldsymbol{\Sigma}^{(\alpha)}$
$\boldsymbol{\tau}^{(\alpha)}$
$\boldsymbol{R}^{(\beta)}$
$\boldsymbol{R}^{(\alpha)}$
$\boldsymbol{\beta}_i$
$\boldsymbol{\alpha}_i$
$\boldsymbol{u}^{(\beta)}$
$\boldsymbol{u}^{(\alpha)}$
$\boldsymbol{X}^{(\beta)}$
d'_{ic}
λ_{id}
$\boldsymbol{X}^{(\alpha)}$
$\eta^{(d')}$
$\eta^{(\lambda)}$
$\theta_{ic}^{(\text{old})}$
$\theta_{id}^{(\text{new})}$
$y_{ic}^{(\text{old})}$
$y_{id}^{(\text{new})}$
$n_{ic}^{(\text{old})}$
$n_{id}^{(\text{new})}$
i individuals

$$
\begin{aligned}
\mu_1^{(\beta)} &\sim \text{Normal}(1, 0.5)\\
\mu_j^{(\beta)} &\sim \text{Normal}(0, 0.5),\ j \in \{2\!:\!6\}\\
\mu_1^{(\alpha)} &\sim \text{Normal}(0.5, 0.5)\\
\mu_2^{(\alpha)} &\sim \text{Normal}(0, 0.5)\\
\tau_j^{(\beta)} &\sim \text{Gamma}(5, 10),\ j \in \{1\!:\!6\}\\
\tau_k^{(\alpha)} &\sim \text{Gamma}(5, 10),\ k \in \{1, 2\}\\
\boldsymbol{R}^{(\beta)} &\sim \text{LKJcorr}(2),\ \boldsymbol{R}^{(\beta)} \in \mathbb{R}^{6\times 6}\\
\boldsymbol{R}^{(\alpha)} &\sim \text{LKJcorr}(2),\ \boldsymbol{R}^{(\alpha)} \in \mathbb{R}^{2\times 2}\\[1ex]
\boldsymbol{\Sigma}^{(\beta)} &= \text{diag}(\boldsymbol{\tau}^{(\beta)}) \cdot \boldsymbol{R}^{(\beta)} \cdot \text{diag}(\boldsymbol{\tau}^{(\beta)})\\
\boldsymbol{\Sigma}^{(\alpha)} &= \text{diag}(\boldsymbol{\tau}^{(\alpha)}) \cdot \boldsymbol{R}^{(\alpha)} \cdot \text{diag}(\boldsymbol{\tau}^{(\alpha)})\\[1ex]
\boldsymbol{\beta}_i &\sim \text{MvNormal}(\boldsymbol{\mu}^{(\beta)},\ \boldsymbol{\Sigma}^{(\beta)}),\ \boldsymbol{\beta}_i \in \mathbb{R}^{6\times 1}\\
\boldsymbol{\alpha}_i &\sim \text{MvNormal}(\boldsymbol{\mu}^{(\alpha)},\ \boldsymbol{\Sigma}^{(\alpha)}),\ \boldsymbol{\alpha}_i \in \mathbb{R}^{2\times 1}\\[1ex]
\omega_j^{(\beta)},\ \omega_k^{(\alpha)} &\sim \text{Uniform}(0.01, 0.3),\ j \in \{1\!:\!6\},\ k \in \{1, 2\}\\
u_{j[study]}^{(\beta)} &\sim \text{Normal}(0, \omega_j^{(\beta)}),\ study \in \{1\!:\!4\}\\
u_{k[study]}^{(\alpha)} &\sim \text{Normal}(0, \omega_k^{(\alpha)}),\ study \in \{1\!:\!4\}\\
\boldsymbol{u}_{[study]}^{(\beta)},\ \boldsymbol{u}_{[study]}^{(\alpha)} &= [u_{1[study]}^{(\beta)}, \ldots, u_{6[study]}^{(\beta)}]^\top,\ [u_{1[study]}^{(\alpha)}, u_{2[study]}^{(\alpha)}]^\top\\[1ex]
c &= \begin{cases} 1, 2, 3 & \text{for phases 1–3}^{(\text{CS}-)}\\ 4, 5, 6 & \text{for phases 1–3}^{(\text{CS}+)} \end{cases}\\
d &= 1 \text{ for CS}-,\ 2 \text{ for CS}+\\
w_c,\ w_d &= \begin{cases} -1/2 & \text{if } c, d \in \{\text{animals}\}\\ \ \ 1/2 & \text{if } c, d \in \{\text{tools}\} \end{cases}\\[1ex]
\eta^{(d')},\ \eta^{(\lambda)} &\sim \text{Normal}(0, 0.5)\\[1ex]
\text{logit}(d'_{ic}) &= \boldsymbol{X}_c^{(\beta)} \cdot (\boldsymbol{\beta}_i + \boldsymbol{u}_{[study]}^{(\beta)}) + w_c\ \eta^{(d')},\ \boldsymbol{X}_c^{(\beta)} \in \mathbb{R}^{1\times 6}\\
\text{logit}(\lambda_{id}) &= \boldsymbol{X}_d^{(\alpha)} \cdot (\boldsymbol{\alpha}_i + \boldsymbol{u}_{[study]}^{(\alpha)}) + w_d\ \eta^{(\lambda)},\ \boldsymbol{X}_d^{(\alpha)} \in \mathbb{R}^{1\times 2}\\[1ex]
\theta_{ic}^{(\text{old})} &= \begin{cases} \Phi(d'_{ic} - \lambda_{i1}), & c \in \{1, 2, 3\}\\ \Phi(d'_{ic} - \lambda_{i2}), & c \in \{4, 5, 6\} \end{cases}\\
\theta_{id}^{(\text{new})} &= \Phi(-\lambda_{id})\\[1ex]
y_{ic}^{(\text{old})} &\sim \text{Binomial}(\theta_{ic}^{(\text{old})}, n_{ic}^{(\text{old})}),\ c \in \{1\!:\!6\}\\
y_{id}^{(\text{new})} &\sim \text{Binomial}(\theta_{id}^{(\text{new})}, n_{id}^{(\text{new})}),\ d \in \{1, 2\}
\end{aligned}
$$

Note. Sensitivity (d') measures the ability to distinguish old from new items, while the criterion (λ) reflects the decision threshold.

SDT assumes two overlapping normal distributions with equal variance: one representing *noise* (new items) and the other representing *signal* (old items). An individual evaluates each item and makes a recognition decision based on whether its memory strength surpasses a predefined criterion.

Two key parameters define performance in this model. Sensitivity (d') measures the ability to discriminate between old (signal) and new (noise) items with higher values indicating better discrimination. Criterion (λ) reflects the decision threshold or bias. It determines how liberal or conservative the participant is when deciding whether an item is *old*. For instance, a liberal criterion yields more hits but also more false alarms. Because there is only one false alarm rate per semantic category in the experimental design, SDT assumes a consistent criterion across all items within that category, regardless of encoding phase. This simplifying assumption, though often reasonable, may conflate true differences in discriminability with unmodeled shifts in response bias. Such shifts may arise if participants adjust caution based on perceived memory strength or source (e.g., being more liberal / conservative for items believed to be shown during conditioning) or base-rate beliefs (e.g., adopting a stricter / weaker criterion if more trials are believed to be shown during conditioning). In such cases, what appears to be a difference in discriminability might instead reflect a change in bias.

A graphical representation of the model, including the assumptions about the data-generating process and the specification of prior distributions, is shown in Figure S4. The overall structure follows the 1HT model, with key modifications in the modeling of hits and false alarms, and in the choice of a few prior distributions. Notably, the model departs from the log-odds space and instead operates in z-units, consistent with the assumption of both the signal and noise distributions to be the standard normal distributions.

The model predicts the number of *hits*, $y_{ic}^{(\text{old})}$, out of $n_{ic}^{(\text{old})}$ trials, with i indexes participants and c indexes the six experimental conditions using a binomial likelihood function with the success probability parameter $\theta_{ic}^{(\text{old})}$, which is defined according to the equation:

$$\theta_{ic}^{(\text{old})} = \begin{cases} \Phi(d'_{ic} - \lambda_{i1}), & \text{when } c \in \text{CS}- \\ \Phi(d'_{ic} - \lambda_{i2}), & \text{when } c \in \text{CS}+, \end{cases}$$

where Φ denotes the cumulative distribution function of the standard normal distribution. The sensitivity parameter d'_{ic} is condition- and participant-specific, while the criterion λ_{id} varies by semantic category (i.e., the semantic category assigned to CS+ vs. CS-) but is held constant across experimental phases (pre-conditioning, conditioning, and post-conditioning).

The hierarchical structure mirrors that of the 1HT and 2HT models. The population-level prior for the grand mean sensitivity parameter across conditions, $\mu_1^{(\beta)}$ was adapted to the z-unit scale and chosen based on prior predictive checks:

$$\mu_1^{(\beta)} \sim \text{Normal}(1, 0.5).$$

The model also predicts the number of *false alarms*, $y_{id}^{(\text{new})}$, out of $n_{id}^{(\text{new})}$ trials, where i indexes participants and d indexes the two semantic categories allocated to CS- and CS+ conditions, using a binomial likelihood function. The success probability parameter $\theta_{id}^{(\text{new})}$ is defined according to the equation:

$$\theta_{i}^{(\text{new})} = \begin{cases} \Phi(-\lambda_{i1}), & \text{for CS-} \\ \Phi(-\lambda_{i2}) & \text{for CS+.} \end{cases}$$

The population-level prior for the grand mean of the bias (i.e., response criterion), $\mu_1^{(\alpha)}$, was also adapted to z-units based on prior predictive checks:

$$\mu_1^{(\alpha)} \sim \text{Normal}(0.5, 0.5).$$

To evaluate model adequacy and potential mis-specification, we conducted posterior predictive checks. These included comparisons between observed and simulated data for summary statistics – specifically, grand means of hits and false alarms – aggregated both across and within studies.

Supplementary Results

Reproducibility checks

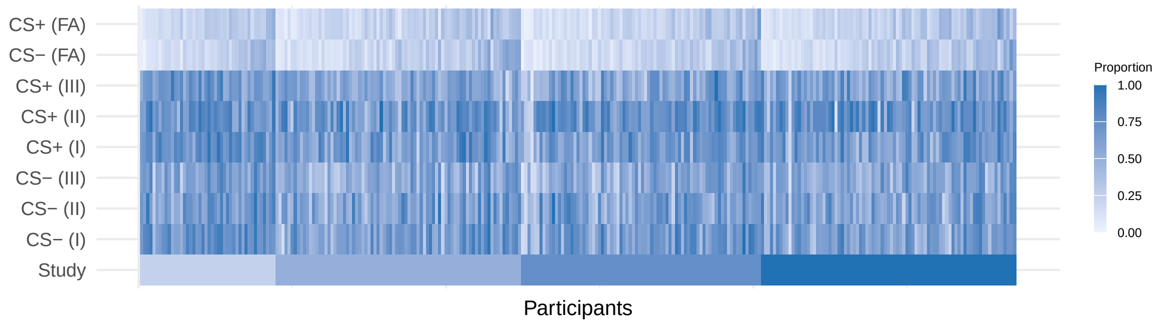
The validity of the retrieved data was successfully verified. Visual inspection revealed that the data are free from errors or inconsistencies (Figure S5). Likewise, we successfully reproduced the main results of the original report (Figure S6).

Latent-mixture two-high-threshold model (2HT): Posterior group membership across studies

To evaluate the consistency of the two-group latent class solution across studies, we examined the representation and separation of the latent groups within each study. The prevalence of the first group was generally consistent across studies, with Study 1 showing the lowest proportion (Study 1: 0.32, Study 2: 0.44, Study 3: 0.49, Study 4: 0.42).

Figure S5

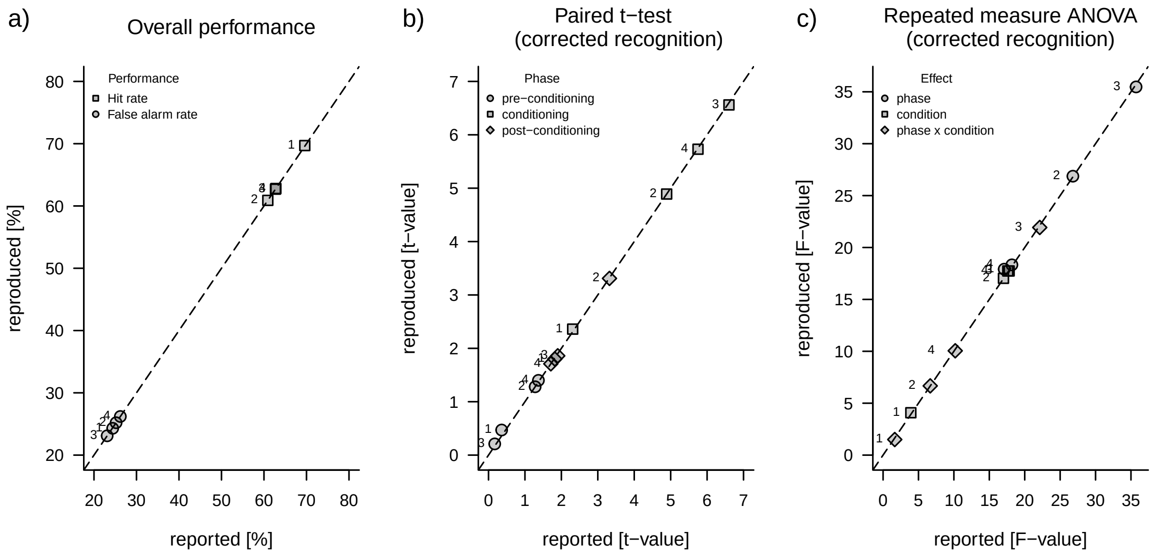
Raw old-new recognition judgments of all participants across four studies.



Note. Raw hit rates and false alarms per participant. Roman numbers indicate phase of the encoding: preconditioning (I), conditioning (II), and post-conditioning (II). Last row indicates the study id that is color-coded.

Figure S6

Reproducibility outcome comparing reported and reproduced values of statistical analyses.



Note. The dashed line is an identity line (x = y). Numbers above the data points indicate study id.

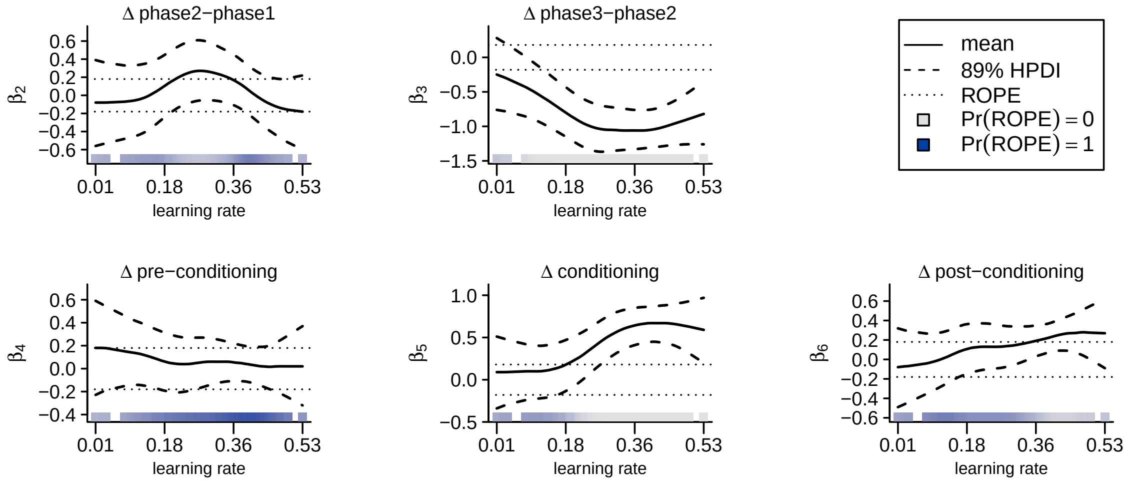
Visual inspection of the ECDFs confirmed the presence of a bimodal distribution in posterior group assignment probabilities across all studies. Overall, the distances between empirical cumulative distribution functions (ECDFs) ranged from small to moderate. Notably, the distribution from Study 1 differed the most from the others (study 1 vs. study 2: KS = 0.19, 89% HPDI [0.11, 0.3]; study 1 vs. study 3: KS = 0.23, 89% HPDI [0.11, 0.34]; study 1 vs. study 4: KS = 0.2, 89% HPDI [0.11, 0.3]). In contrast, the distributions from Study 2 and Study 4 were highly similar, with minimal divergence (KS = 0.11, 89% HPDI [0.06, 0.16]).

These findings suggest that, despite some variability across studies – most notably in Study 1, likely due to procedural differences – the inferred group structure is largely consistent and not driven solely by study-specific factors. In other words, both latent groups are represented in each study, even if their relative proportions vary.

Gaussian Process extension of two-high-threshold model (2HT)

Figure S7

Results of the Gaussian Process extension of two-high-threshold model (2HT) considering learning rate.



Note. Plots depict latent functions that map input features (learning rate) onto $\boldsymbol{\beta}_i$ parameters (e.g., CS+ vs. CS- difference nested within phases). Values of learning rate are bounded between 0 and 1. $\mathrm{Pr}_{\mathrm{ROPE}}$ indicates the proportion of the posterior at a particular value of the learning rate that falls within the region of practical equivalence, which is +/-0.18 on the log-odds scale.

We explored whether individual differences in Pavlovian learning contribute to recognition memory performance using the 2HT model extended with a Gaussian Process (GP).

Learning rate

First, we considered the role of the α learning rate of the RW model (Figure S7), which reflects how quickly individuals update the associative strength between a semantic category and the electric shock. This metric essentially serves as an index of how effectively participants distinguish between the two categories.

The estimated length scale parameter was relatively high (l = 3.28, 89% HPDI [1.63, 5.89]), indicating that the underlying functions were fairly smooth, with minimal variation in response to small changes in the learning rate. The GP accounted for a moderate level of noise in the participant-level recognition parameters (σ_{gp} = 0.6, 89% HPDI [0.56, 0.64]), suggesting the presence of some process noise.

When examining individual profiles of GP latent functions, we observed a striking divergence in selective memory prioritization for items encoded during conditioning (CS+ vs. CS-) between participants with low vs. high learning rates. Specifically, individuals who rapidly learned the category-shock association exhibited a clear selective memory advantage for CS+ items over CS- items (i.e., the latent function lies outside the ROPE). In contrast, for individuals with slower learning rates, this effect was unreliable, with the latent function primarily overlapping with ROPE.

A similar pattern emerged when comparing recognition memory across encoding phases. Individuals with faster learning rates showed better recognition for items encoded during conditioning – regardless of semantic category – paired with lower recognition for items encoded post-conditioning. The reverse was true for slower learners.

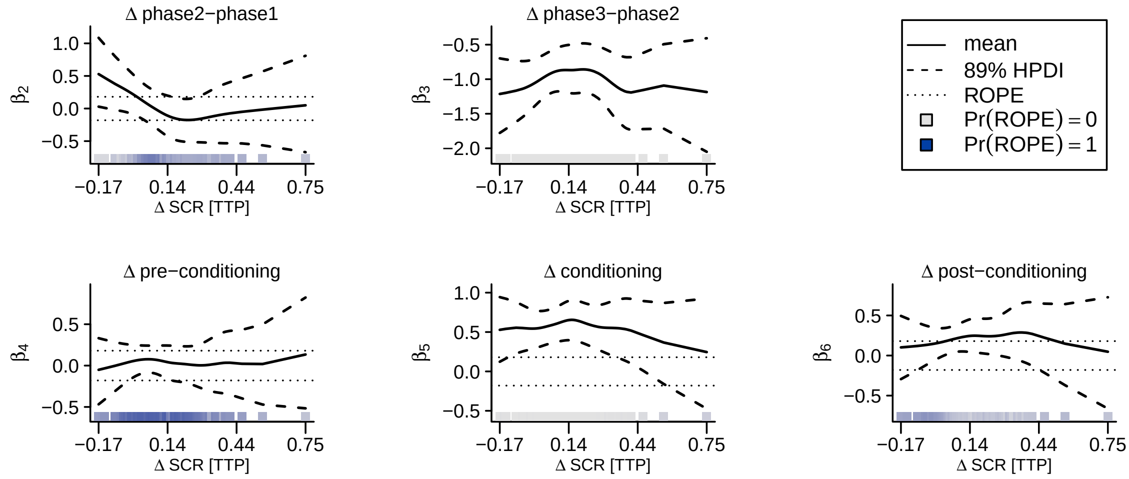
Additionally, we observed a qualitative linear trend in selective memory prioritization for items encoded during post-conditioning as a function of learning rate: as the learning rate increased, the latent function progressively moved away from ROPE. However, due to high uncertainty in the estimates, these observations remain inconclusive.

Overall, beyond these specific effects, the learning rate showed limited relevance for explaining broader individual differences in recognition memory performance.

Differential anticipatory SCRs

Figure S8

Results of the Gaussian Process extension of two-high-threshold model (2HT) considering differential anticipatory SCR.



Note. Plots depict latent functions that map input features (CS+ vs. CS- difference in anticipatory SCR estimated with through-to-peak (TTP) method) onto $\boldsymbol{\beta}_i$ parameters (e.g., CS+ vs. CS- difference nested within phases). Values of ΔSCR are unbounded. Pr_{ROPE} indicates the proportion of the posterior at a particular value of the ΔSCR that falls within the region of practical equivalence, which is +/-0.18 on the log-odds scale.

The estimated length scale parameter was very high (l = 6.43, 89% HPDI [3.19, 12.26]), indicating that the underlying functions were very smooth, or perhaps even resembling a straight line with minimal variation in response to big changes over input space. Similarly to the GP with learning rate, the model accounted for a moderate level of noise in the participant-level recognition parameters (σ_{gp} = 0.59, 89% HPDI [0.55, 0.63]), suggesting the presence of some process noise.

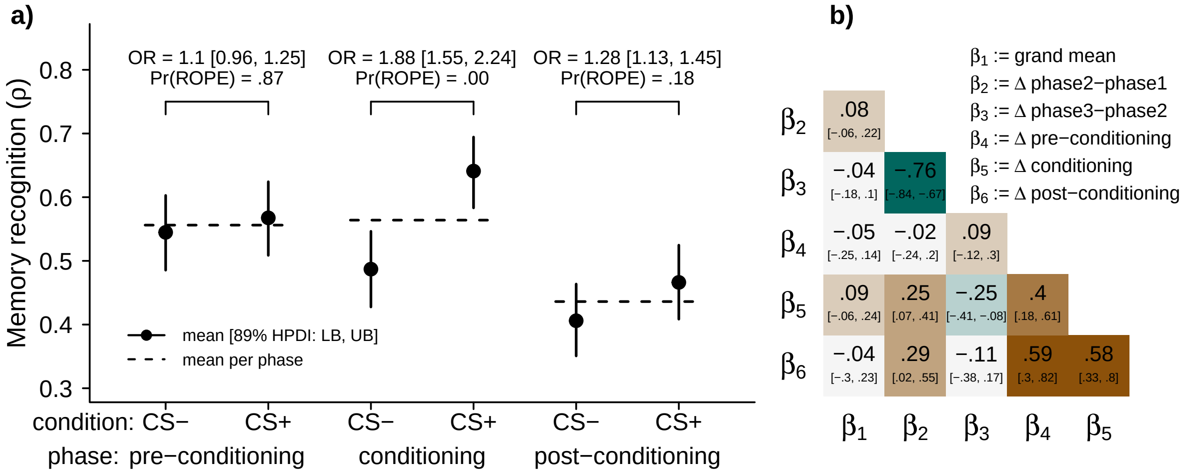
When considering the individual profiles of latent functions over the differential anticipatory SCRs, we observed no clear pattern except the category-selective memory prioritization effect for items encoded during conditioning (Figure S8). Medium-size difference in SCR corresponded with a clear effect (i.e., the latent function outside ROPE). These individuals also showed a reduced uncertainty in the putative effect of selective memory prioritization for items encoded in the post-conditioning phase.

Together, SCR proved to be only partially relevant in explaining selective memory prioritization for items encoded during conditioning, and it could not account for why the effect might appear in the pre- or post-conditioning phases.

Results of the one-high-threshold model (1HT)

Figure S9

Results of the Bayesian 1HT model of old-new recognition judgments.



Note. Panel a depicts posterior estimates of memory recognition (ρ). Square brackets indicate 89% HPDI interval of odds ratio (OR). Proportion of the posterior falling inside the region of practical equivalence is denoted by Pr(ROPE). Panel b depicts within-subject correlation matrix of random effects (mean [89% HPDI: LB, UB]).

The results of the Bayesian 1HT model largely mirrored those of the 2HT model (Figure S9), with only minor discrepancies – particularly slightly higher raw recognition estimates and stronger participant-level correlations of random effects. These differences likely stem from how the two models handle false alarms: in the 1HT model, false alarms arise solely from guessing.

Overall, the 1HT model produced slightly higher raw recognition estimates (~0.52 vs. ~0.49 in the 2HT model). Similar to the 2HT results, we observed a clear memory prioritization effect for CS+ items encoded during conditioning (OR = 1.88, 89% HPDI [1.55, 2.24], $\text{Pr}_{\text{ROPE}} = .00$), with higher recognition for CS+ ($\rho = .64$, 89% HPDI [.58, .69]) than CS- items ($\rho = .49$, 89% HPDI [.43, .55]). Recognition rates for items encoded during pre-conditioning were comparable across categories ($\rho_{CS+} = .57$, 89% HPDI [.51, .62]; $\rho_{CS-} = .54$, 89% HPDI [.49, .6]; OR = 1.1, 89% HPDI [0.96, 1.25], $\text{Pr}_{\text{ROPE}} = .87$), whereas post-conditioning recognition was slightly higher for CS+ items ($\rho_{CS+} = .47$, 89% HPDI [.41, .52]; $\rho_{CS-} = .41$, 89% HPDI [.35, .46]; OR = 1.28, 89% HPDI [1.13, 1.45], $\text{Pr}_{\text{ROPE}} = .18$), suggesting tentative evidence for selective and proactive memory prioritization – consistent with the 2HT model's findings.

The overall correlation structure among participant-level random effects was similar to that in the 2HT model, though somewhat stronger. Notably, we found robust correlations in selective memory prioritization across phases: pre-conditioning and post-conditioning (r = .59, 89% HPDI [.3, .82], $\text{Pr}_{\text{ROPE}} = .01$), conditioning and post-conditioning (r = .58, 89% HPDI [.33, .8], $\text{Pr}_{\text{ROPE}} = .00$), as well as, pre-conditioning and conditioning (r = .4, 89% HPDI [.18, .61], $\text{Pr}_{\text{ROPE}} = .02$). These results suggest that – more so than in the 2HT model – some individuals consistently prioritized memory for CS+ items across all encoding phases, relative to the group average. However, these correlations should be interpreted with caution, as they may be confounded by the fact that items across all three phases belonged to the same semantic category. In other words, participants with better memory for one category (e.g., "tools" over "animals") may show consistently higher performance across all three encoding phases in the condition to which that category was assigned (and vice versa).

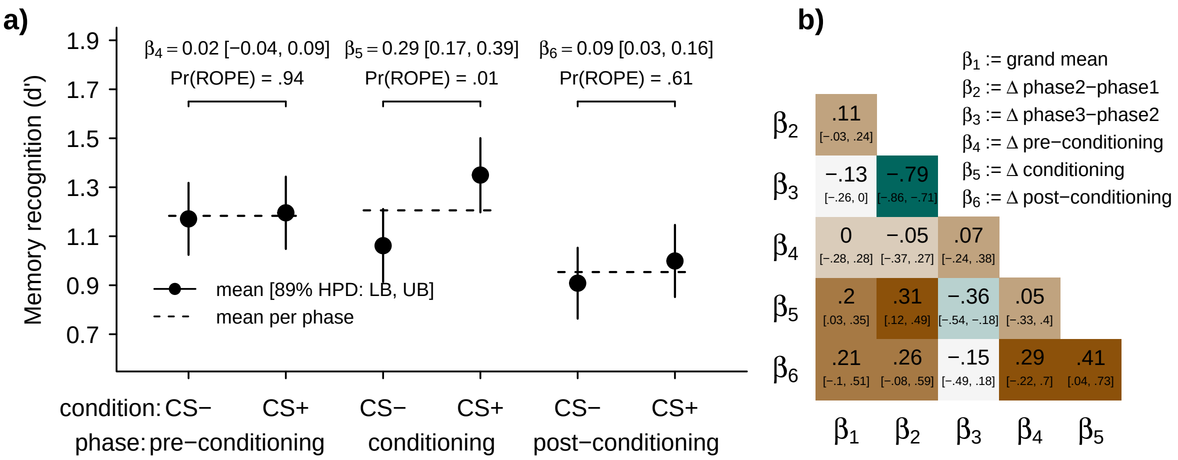
In contrast to the 1HT model, the 2HT model showed weaker participant-level correlations between phase-specific recognition parameters. This difference arises because, in the 2HT model, all recognition parameters jointly contribute to a single false alarm rate, introducing statistical dependence among them. This coupling acts as a form of regularization, constraining recognition estimates across phases to jointly account for false alarm behavior. In the 1HT model, false alarms are modeled independently of recognition, allowing phase-specific parameters to vary more freely. This structural difference likely accounts for the stronger cross-phase correlations observed in the 1HT model.

Consequently, the 2HT model may better mitigate confounding influences – such as participants having generally better memory for a particular semantic category assigned to CS+ – even though it exhibits slightly lower predictive performance. In contrast, the 1HT model reveals stronger participant-level correlations, which may reflect genuine individual differences but also greater sensitivity to such confounds. By constraining recognition through a shared false alarm rate, the 2HT model provides a more conservative and potentially more robust estimate of cross-phase consistency, albeit at the cost of reduced flexibility.

Results of model based on signal detection theory (SDT)

Figure S10

Results of the Bayesian SDT model of old-new recognition judgments.



Note. Panel a depicts posterior estimates of sensitivity (d'). Square brackets indicate 89% HPDI interval. Proportion of the posterior falling inside the region of practical equivalence is denoted by Pr(ROPE), which is -0.1 and 0.1. Panel b depicts within-subject correlation matrix of random effects (mean [89% HPDI: LB, UB]).

The results of the Bayesian SDT model largely mirrored the qualitative pattern observed with the 2HT model (Figure S10), although the effect sizes were somewhat attenuated. To determine the presence of an effect in the sensitivity estimates (d'), we used a ROPE with a range of [-0.1, 0.1].

As in the other two models, we observed a clear memory prioritization effect for CS+ items encoded during conditioning ($\Delta\beta$ = 0.29, 89% HPDI [0.17, 0.39], Pr$_{\text{ROPE}}$ = .01), with higher recognition for CS+ (d' = 1.35, 89% HPDI [1.2, 1.5]) than CS- items (d' = 1.06, 89% HPDI [.91, 1.21]). Sensitivity values for items encoded during pre-conditioning were comparable across categories (d'_{CS+} = 1.2, 89% HPDI [1.05, 1.34]; d'_{CS-} = 1.17, 89% HPDI [1.02, 1.32]; $\Delta\beta$ = 0.02, 89% HPDI [-0.04, 0.09], Pr$_{\text{ROPE}}$ = .94), and post-conditioning recognition was slightly higher for CS+ items (d'_{CS+} = 1, 89% HPDI [.85, 1.15]; d'_{CS-} = .91, 89% HPDI [.76, 1.05]; $\Delta\beta$ = 0.09, 89% HPDI [0.03, 0.16], Pr$_{\text{ROPE}}$ = .61), but this difference largely fell within the ROPE, tentatively suggesting no meaningful effect. Although the post-conditioning effect had a lower probability of being distinct from zero than in the other models, it still showed a clearly positive direction.

Participant-level correlations between phase-specific memory estimates were generally weaker in the SDT model than in the 1HT model, despite both models sharing similar structural assumptions, i.e., a single parameter contributing to false alarm rates (criterion λ). Notably, the correlation of participant-level random effects between the pre-conditioning and conditioning phases was negligible (r = .05, 89% HPDI [-0.33, .4], Pr$_{\text{ROPE}}$ = .32), indicating little to no consistent relationship across participants. The correlation between pre-conditioning and post-conditioning phases was somewhat stronger (r = .29), though accompanied by considerable uncertainty (89% HPDI [-0.22, .7], Pr$_{\text{ROPE}}$ = .14). In contrast, the correlation between conditioning and post-conditioning phases was moderate and more robust (r = .41, 89% HPDI [.04, .73], Pr$_{\text{ROPE}}$ = .06), suggesting a more stable relationship in participant-level estimates across these phases. Finally, we observed a strong negative correlation between phase-to-phase differences, regardless of semantic category (r = -0.79, 89% HPDI [-0.86, -0.71], Pr$_{\text{ROPE}}$ = .00) that was also present in other two models.

Despite their structural similarity, the SDT and 1HT models yielded different participant-level correlations across phases. The dampened correlations likely arise from how each model handles memory evidence and shared parameters. The SDT model uses a shared decision criterion (λ) to transform phase-specific sensitivity estimates (d') into hit probabilities via a nonlinear (Φ) function, i.e., each phase-specific sensitivity is scaled by a shared decision criterion. The shared criterion introduces dependency across phases and reduces independent variability in sensitivity estimates. Moreover, the nonlinearity of the cumulative distribution function (CDF) of the standard normal distribution (Φ) compresses differences in the link space – particularly near the extremes – further contributing to the dampening of correlations. In contrast, the 1HT model expresses recognition as a linear mixture of phase-specific recognition and a shared guessing parameter, allowing phase-specific recognition estimates to vary more freely. As a result, individual differences of recognition in 1HT are more directly expressed, leading to higher correlations across phases. In simple terms: the 1HT model allocates nearly all variation to recognition, while the SDT model distributes it between sensitivity and criterion, blurring phase-specific signal estimates.

References

Kalbe, F., & Schwabe, L. (2021). On the search for a selective and retroactive strengthening of memory: Is there evidence for category-specific behavioral tagging? *Journal of Experimental Psychology. General*. https://doi.org/10.1037/xge0001075

Rescorla, R. A., & Wagner, A. R. (1972). A theory of Pavlovian conditioning: Variations on the effectiveness of reinforcement and non-reinforcement. In A. H. Black & W. F. Prokasy (Eds.), *Classical conditioning II: Current research and theory* (pp. 64–99). Appleton-Century-Crofts.

Stan Development Team. (2023). *Stan modeling language users guide and reference manual*. Stan Development Team. https://mc-stan.org

Stan Development Team. (2024). *RStan: The R interface to Stan*. https://mc-stan.org/

Tzovara, A., Korn, C. W., & Bach, D. R. (2018). Human pavlovian fear conditioning conforms to probabilistic learning. *PLOS Computational Biology*, *14*(8), e1006243. https://doi.org/10.1371/journal.pcbi.1006243

Vermunt, J. K. (2010). Latent class modeling with covariates: Two improved three-step approaches.

Political Analysis, *18*(4), 450–469. http://www.jstor.org/stable/25792024