Supplemental Material

Kevin P. Darby, Per B. Sederberg
Department of Psychology, University of Virginia

Regression models

To assess performance statistically, we performed a series of hierarchical Bayesian regression models, which estimated effects for each age group overall while accounting for participant-level effects. To analyze choices, we predicted "old" responses with logistic regressions, whereas we predicted RTs with linear regressions. For each analysis, we estimated intercept and coefficient terms for each age group separately, but did not include any interaction terms in order to simplify the models. Each Bayesian analysis resulted in a posterior distribution for each term in the regression, for each age group. To assess whether independent variables had an effect, we determined whether the 95% highest posterior density (HPD) of the posterior distribution of each regression slope contained zero. As a more continuous measure we also determined the percentage of samples in each posterior that were above (or below) zero, which should approach 100% for strong effects. To assess age differences, we calculated $\hat{\eta}$, a measure of overlap between two distributions (Pastore & Calcagnì, 2019), as described in the Results section of the main text. We fit each regression using RunDEMC (https://github.com/compmem/RunDEMC), a Python library implementing Bayesian differential evolution Markov chain Monte Carlo techniques [DEMC; Turner, Sederberg, Brown, & Steyvers (2013)].

Every regression parameter θ (both intercepts and coefficients), for every participant i in age group j, was modeled with a normal prior distribution:

$$\theta_{i,j} \sim \mathcal{N}(\theta_{\mu_j}, \theta_{\sigma_j}),$$

where θ_{μ_j}, and θ_{σ_j} are group-level hyper-parameters controlling the mean and standard deviation, respectively, of the participant-level prior distributions. These hyper-parameters were modeled with the following hyper-priors:

$$\theta_{\mu_j} \sim \mathcal{N}(0, 1)$$

$$\theta_{\sigma_j} \sim InvGamma(1, 1).$$

We fit each regression model with 2500 iterations of the MCMC algorithm, the first 500 of which consisted of a burnin period. This process occurred in 100 chains, and we included the final 500 iterations of every chain in the analysis.

The first statistical model we performed assessed discrimination between old and new pairs. To do so, we compared hits (i.e., correct "old" responses) for repeated intact pairs to false alarms (i.e., incorrect "old" responses) to both New and Recombined pairs, separately and averaged across strength conditions, by applying a regression model with a logistic linking function. Prior work has demonstrated that this regression approach is equivalent to the classical d' index of signal detection theory based on the cumulative distribution function of the Gaussian distribution, although the logistic linking function may cause the values to be scaled differently (DeCarlo, 1998).

The intercept of this analysis was lower in older adults, $95\% HPD_{older}[1.22, 1.53]$, compared to young adults, $95\% HPD_{young}[1.89, 2.13]$. This difference was credible, $\hat{\eta} < .001$, indicating fewer hits to intact pairs in older adults. The posterior distributions were credibly above zero for the $d_{recombined}$ values for both young adults, $95\% HPD_{young}[3.25, 3.72]$, 100% above zero, and older adults, $95\% HPD_{older}[1.64, 2.13]$, 100% above zero, indicating discrimination between intact and recombined pairs. However, older adults had lower $d_{recombined}$ values compared to young adults, $\hat{\eta} < .001$, which supports prior work suggesting that older adults are not as able as young adults to remember associations between items (Naveh-Benjamin, 2000). This analysis also assessed young and older adults' ability to discriminate between New and repeated Intact pairs. d_{new} values were reliably above zero for young adults, $95\% HPD_{young}[5.79, 6.48]$, 100% above zero, as well as older adults, $95\% HPD_{older}[4.60, 5.52]$, 100% above zero. Older adults had lower d_{new} values, $\hat{\eta} = .011$, indicating that older adults were less able to discriminate between pairs with new items and repeated intact pairs.

We also examined whether RTs (in seconds), which were added to 1 and log-transformed, differed between New or Recombined pairs in comparison to repeated Intact pairs using a separate hierarchical Bayesian multiple regression model. The intercept of this analysis was higher for older adults, $95\% HPD_{older}[0.79, 0.90]$, compared to young adults, $95\% HPD_{young}[0.64, 0.72]$, and this difference was credible, $\hat{\eta} = .001$, indicating slower responses overall in older adults. In addition, RTs were generally faster in New compared to repeated Intact pairs for both young adults, $95\% HPD_{young}[-0.10, -0.03]$, >99.9% below zero, and older adults, $95\% HPD_{older}[-0.11, .01]$, 95.7% below zero, but we found little evidence that this effect was different between age groups, $\hat{\eta} = .68$. In addition, RTs were slower for recombined pairs compared to repeated intact pairs, in both young adults,

$95\%HPD_{young}[0.05, 0.12]$, 100% above zero, and older adults, $95\%HPD_{older}[0.01, 0.12]$, 98.9% above zero, with no difference between age groups ($\hat{\eta}$ = .67).

We also analyzed factors impacting false alarms to Recombined pairs and hits to Intact pairs separately. To assess false alarms to Recombined pairs, we conducted a hierarchical Bayesian logistic regression predicting false alarms depending on the categorical "strength" condition (i.e., whether the items had been seen in their original pairings once, twice, or three times prior to being recombined). The intercept of this analysis was higher for older adults ($95\%HPD_{older}[-0.65, -.20]$) compared to young adults ($95\%HPD_{young}[-1.31, -0.95]$), $\hat{\eta}$ = .001, indicating that older adults were generally more likely to false alarm to recombined pairs. Greater strength was predictive of lower false alarms in young adults, $95\%HPD_{young}[-0.59, -0.28]$, 100% below zero, but less so in older adults, $95\%HPD_{older}[-0.30, 0.06]$, 91.0% below zero. The difference between age groups was not very robust, $\hat{\eta}$ = .068, although the trend is consistent with prior work (Gallo, Sullivan, Daffner, Schacter, & Budson, 2004; Light, Patterson, Chung, & Healy, 2004) suggesting that young adults, but not older adults, may be able to form stronger associative memory for pairs that are repeated, and as a result be better able to reject subsequently recombined pairs.

We also examined whether the strength condition affected RTs to Recombined pairs. The intercept of this regression was higher for older adults ($95\%HPD_{older}[0.86, 0.98]$) compared to young adults ($95\%HPD_{young}[0.74, 0.83]$), $\hat{\eta}$ = .007, indicating overall slower RTs to Recombined pairs in older adults. We did not find very strong evidence that higher strengths predicted faster RTs in young adults, $95\%HPD_{young}[-0.06, 0.01]$, 89.2% below zero, or in older adults, $95\%HPD_{older}[-0.07, 0.05]$, 65.7% below zero, and the difference between distributions was not robust, $\hat{\eta}$ = .719.

We also analyzed hits to Intact pairs. Specifically, we assessed repetition effects, as well as potential interference effects on memory for Intact pairs following the presentation of Recombined pairs. To do so, we coded whether a pair had previously been presented once (i.e., Intact 1 pairs) or twice (i.e., Intact 2 pairs). We expected that hit rates would be higher in both age groups for pairs that had already been presented twice (Light, Patterson, Chung, & Healy, 2004)). In addition, we coded whether intact pairs had been recombined on the immediately preceding presentation of the items (i.e., Weak Intact 1 and Medium Intact 2 pairs). We hypothesized that these pairs may be more difficult to remember due to interference from the recombined pairs.

We applied another hierarchical Bayesian regression to investigate these issues. The intercept of this analysis was lower in older adults ($95\%HPD_{older}[0.93, 1.35]$) compared to young adults ($95\%HPD_{young}[1.74, 2.10]$), $\hat{\eta}$ < .001, indicating overall reduced accuracy for Intact pairs. In addition, we found strong evidence of higher hit rates following two previous presentations in both young adults, $95\%HPD_{young}[1.05, 1.43]$, 100% above zero, and older adults, $95\%HPD_{older}[0.83, 1.30]$, 100% above zero, with no evidence of age differences, $\hat{\eta}$ = .424. We also found evidence of lower hit rates following Recombined pairs in both age groups: young adults $95\%HPD_{young}[-0.93, -0.56]$, 100% below zero, and older adults $95\%HPD_{older}[-0.63, -0.19]$, >99.9% below zero. Although this likely interference effect tended to be somewhat stronger in young adults, there was not strong evidence of a difference between distributions, $\hat{\eta}$ = .099.

We also examined whether repetition and interference affected RTs to intact pairs. The intercept of this analysis, $95\%HPD_{young}[0.65, 0.73]$ for young adults and $95\%HPD_{older}[0.81, 0.92]$ for older adults, indicated slower RTs to Intact pairs in older adults, $\hat{\eta}$ < .001. In addition, we found evidence of faster RTs with an additional repetition in young adults, $95\%HPD_{young}[-0.11, -0.04]$, >99.9% below zero, and older adults, $95\%HPD_{older}[-0.12, -0.01]$, 98.4% below zero, with no difference between age groups, $\hat{\eta}$ = .749. We also found evidence of slower RTs following Recombined pairs in young adults, $95\%HPD_{young}[0.05, 0.12]$, 100% above zero. We found less evidence of this effect in older adults, $95\%HPD_{older}[-0.02, 0.09]$, 88.7% above zero, although we did not find strong evidence of a difference between age groups, $\hat{\eta}$ = .262.

Computational model description

We begin by describing the decision-making component of the model. To simulate choices and RTs, we pass memory strengths estimated on every trial to a Wiener first passage of time model (Navarro & Fuss, 2009; Stone, 1960), a type of sequential sampling model (Ratcliff & McKoon, 2008). In this model, evidence noisily accumulates across time until it crosses one of two response thresholds: one for an "old" response, estimated by the free parameter a, or one for a "new" response, which was set at zero. The starting point of this evidence accumulation, which could be biased toward the "new" or "old" responses, was also estimated by a free parameter, w. A w value of 0.5 corresponds to an unbiased starting point directly between the two response options, whereas a value less than 0.5 biases the decision toward "new," and a value greater than 0.5 biases the decision toward "old." Evidence accumulates noisily, driven by a "drift rate," representing the quality of evidence for one or the other decision, which is determined by the difference between the memory strength supporting an "old" response and the strength supporting a "new" response. Given the drift rate, as well as the decision threshold a and bias w, we simulated the response and decision time of each trial. Note that this decision time was added to a non-decision time, t_0, which estimates the duration of motor and perceptual processes unrelated to the decision itself.

We will now describe the memory portion of the model, which determines the memory strengths that drive the decision-making process for every trial. Each object presented in the CAR task is represented in the model as a vector of features, which is orthogonal to the representation of every other object. Specifically, each object is represented distinctly as a vector $\boldsymbol{f}$ with maximal activation of one unique feature (i.e. a value of 1.0), and zero activation of all other features. Temporal context is instantiated as a separate vector $\boldsymbol{t}$, the features of which become activated as items are presented, as described below (see Fig. S1). The context vector $\boldsymbol{t}$ represents the same features as the items, such that the context is primarily composed of item features, and changes as different items are presented. The exception to this is that to initialize context, we activate an additional pre-experimental context feature prior to presenting any experimental items to context. This feature represents pre-experiment experiences, such as arriving at the laboratory and the participant's mood before beginning the task. Objects are associated with the state of temporal context at the time when they are presented by means of prediction-error learning; these associations are stored in matrix $\boldsymbol{M}$.

Familiarity is estimated in the model by first determining the activation of the two items within the current state of temporal context, $\boldsymbol{t_{i-1}}$, which was last updated for the previous trial:

$$s_{ro} = (\boldsymbol{f_A} + \boldsymbol{f_B}) \cdot \boldsymbol{t_{i-1}}.$$

The $\cdot$ symbol represents a dot product between the two item representations ($\boldsymbol{f_A} + \boldsymbol{f_B}$) and current context ($\boldsymbol{t_{i-1}}$); this operation simply reads out the activation of the two item features as a single value, s_{ro}. This readout of activation is then converted to a strength of familiarity via the following exponential function:

$$s_f = \lambda(1 - e^{\frac{-s_{ro}}{\tau}}).$$

This exponential function creates a non-linear mapping from the readout signal onto a scale from 0 to a maximum value of free parameter λ (see Fig. S2). As the readout strength grows larger, the resulting value approaches λ, and the steepness of this function is governed by the free parameter τ, such that as τ approaches zero, the function becomes step-like, such that any activation of the item features will result in near-maximal familiarity strength, whereas with higher values of *tau* the function is more shallow. As the activation of items in context decays as new items are presented (as in Fig. S1), τ helps modulate how quickly familiarity fades across time. By contrast, λ helps control the maximum or asymptotic magnitude of familiarity strength. Note that items in New pairs have no activation in context, and therefore $s_f = 0$, regardless of the values of λ and τ.

In addition to this familiarity signal, the model also instantiates associative memory by reinstating the temporal contexts in which each item was presented, and calculating the match (or overlap) between the reinstated contexts, as well as the mismatch (or difference) between them. New items have not been associated with any contexts and result in no match or mismatch. Items presented previously only in the same pair will reinstate identical contexts, producing a strong match signal and no mismatch signal. Items presented in pairs seen in similar but not identical contexts (i.e., close together in the list of pairs but not in the same pair) would be expected to produce a somewhat strong match signal and a somewhat weak mismatch signal, whereas items presented in pairs from far apart in the list would be expected to produce a weak match signal and strong mismatch signal.

To reinstate the contexts associated with each item in a pair, we simply take a dot product between each of the items, $\boldsymbol{f_A}$ and $\boldsymbol{f_B}$, and the associative matrix $\boldsymbol{M}$:

$$\boldsymbol{t'_A} = \boldsymbol{f_A} \cdot \boldsymbol{M}$$

$$\boldsymbol{t'_B} = \boldsymbol{f_B} \cdot \boldsymbol{M}.$$

This results in two vectors representing the contexts previously associated with each item. To calculate the match between

$$\boldsymbol{t'_A}$$

and t'_B, we take the dot product between these vectors:

$$s_m = \boldsymbol{t'_A} \cdot \boldsymbol{t'_B},$$

which results in a single number corresponding to the overlap between the context vectors (s_m).

Similarly, we calculate the *mismatch* between the two reinstated context vectors by taking the difference between them, and then taking the dot product of this mismatch signal with itself (which is equivalent to taking the sum of squared distance):

$$s_{mm} = \gamma\sqrt{(\boldsymbol{t'_A} - \boldsymbol{t'_B}) \cdot (\boldsymbol{t'_A} - \boldsymbol{t'_B})}.$$

The square root of the mismatch strength (s_{mm}) is taken to counteract the fact that the dot product takes every retrieved context difference into account twice, and the free parameter γ allows for individual differences in sensitivity to the mismatch of retrieved contexts.

Once these memory strengths are calculated, they are combined into a single value: $s = s_f + s_m - s_{mm} - \nu$, where ν estimates the novelty-driven baseline strength for "new" responses. This s value determines the balance of evidence for "new" and "old" responses and is used as the drift rate that drives evidence accumulation in the decision-making portion of the model to simulate choices and RTs, as described above. Once the memory strengths have been calculated, new learning is allowed to occur. Learning is based on prediction error, where positive prediction error reflects the presence of the unexpected, whereas negative prediction error reflects absence of the expected. For each trial, the model predicts what items are expected to be presented based on the current

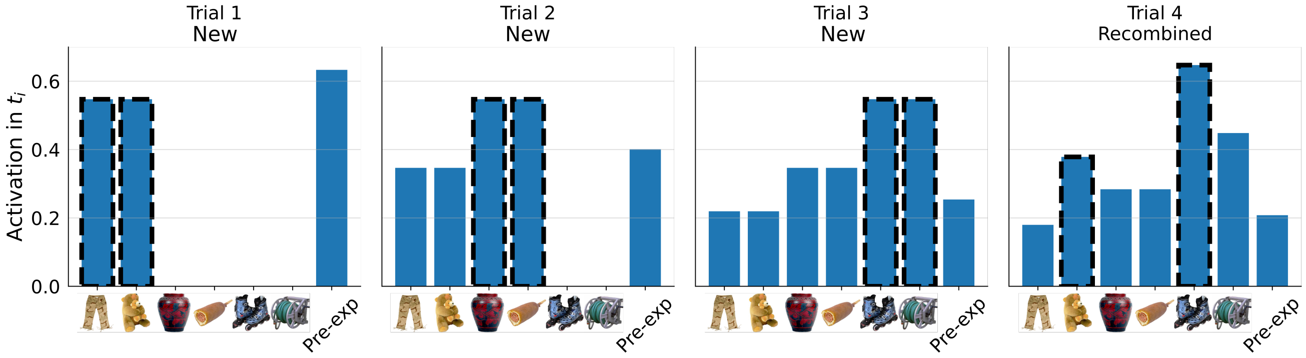

*Figure S1. Activation of context features in context **t** on trial i as pairs of items are presented in the CAR task.* The feature activations corresponding to the presented items are highlighted with dashed rectangles. Contextual features correspond to different items, except for the pre-experimental context unit, labeled as “Pre-exp.”

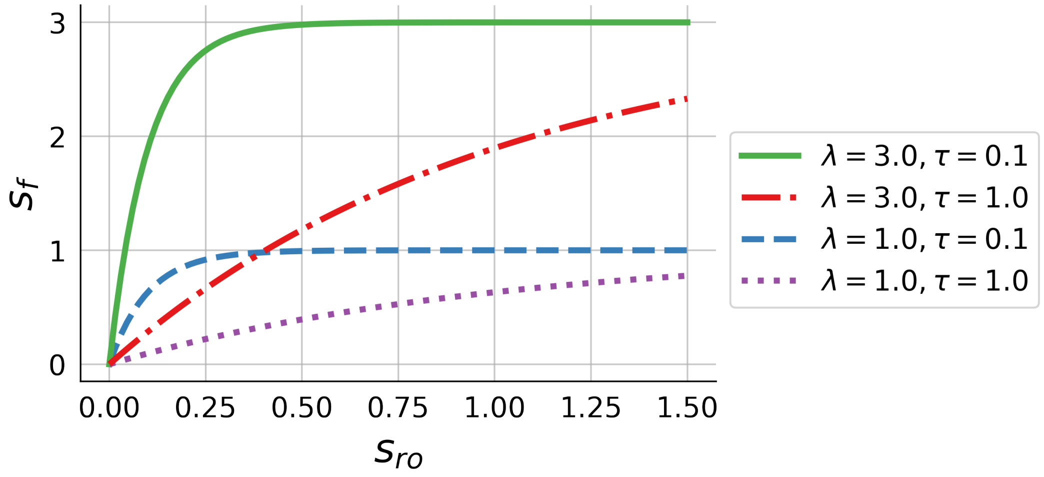

Figure S2. The transformation from the summed activation of items in context (s_{ro}) to familiarity strength (s_f). The transformation is controlled by parameters controlling the maximum s_f value (λ), and the steepness of the function (τ).

state of context, which has not been updated with the currently presented items:

$$\boldsymbol{f}' = \boldsymbol{M} \cdot \boldsymbol{t_{i-1}}$$

This results in an array of item activation values, which we constrained to be between 0 and 1, since the activation of item representations are either 0 or 1. We then compare the presented items with the predicted item activations by taking the differences between them:

$$\boldsymbol{f}'_{\boldsymbol{e}} = \boldsymbol{f} - \boldsymbol{f}'$$

where $\boldsymbol{f} = \boldsymbol{f_A} + \boldsymbol{f_B}$. This results in a vector, in which positive values correspond to positive prediction errors and negative values correspond to negative prediction errors. We then separately bind positive prediction errors ($\boldsymbol{f}'_{\boldsymbol{e^+}}$) and negative prediction errors ($\boldsymbol{f}'_{\boldsymbol{e^-}}$) to the current state of temporal context with outer products that are added to the matrix $\boldsymbol{M}$:

$$\boldsymbol{M_i} = \boldsymbol{M_{i-1}} + \alpha(\boldsymbol{f}'_{\boldsymbol{e^+}} \bigotimes \boldsymbol{t_{i-1}} + \kappa \boldsymbol{f}'_{\boldsymbol{e^-}} \bigotimes \boldsymbol{t_{i-1}}).$$

In this equation, α controls the magnitude of associative learning, and is estimated for each participant as a free parameter. Note that due to this equation, items that are new or have not been presented for some time will be surprising, and as a result they will be strongly bound to the current state of context. At the same time, items that the model incorrectly expected to be presented based on current context, are *unbound* from context by weakening those associations. This unbinding or unlearning is scaled by an additional parameter, κ. If the parameter κ is equal to 1, positive and negative prediction error learning are symmetric; if κ is zero, no unlearning takes place, and if κ is greater than 1, more unlearning takes place than learning. This prediction error mechanism is in contrast to most implementations of TCM, in which once-presented items are simply bound to the current state of context via a Hebbian association (e.g., $\boldsymbol{M_i} = \boldsymbol{M_{i-1}} + \alpha((\boldsymbol{f_A} + \boldsymbol{f_B}) \bigotimes \boldsymbol{t_{i-1}})$).

In the final stage of processing for each trial, temporal context is updated according to the following equation:

$$\boldsymbol{t_i} = |(1 - r\rho)\boldsymbol{t_{i-1}} + r\rho \boldsymbol{t^{IN}}|.$$

The left side of this equation results in decay of the features in the current context, and the right side updates context with the new input, $\boldsymbol{t^{IN}}$, as described below. Values of ρ close to zero would result in very little change of context across trials, such that new input would have relatively little effect on context, whereas when ρ is closer to 1, new items replace items already in context at a faster rate. The |.| symbols indicate that the updated context vector is normalized to unit length. The purpose of this normalization is so that the magnitude of overall activation within context does not vary greatly between trials as more features become activated.

The value r is a sigmoidal novelty signal, described below, that ranges between 0 and 1. As items are presented multiple times in this experiment, we assume that repeated items are not encoded into context as strongly as new items, and that context does not change as rapidly when an item is repeated compared to when it is new. This mechanism has been previously applied to TCM in situations with repeated item features (Siefke, Smith, & Sederberg, 2019). Following Siefke et al. (2019), we first calculate a term r from an exponential function that estimates novelty due to the stimulus readout from context for each trial:

$$r = e^{-s_{ro}}.$$

For new items, $s_{ro} = 0$, and $r = 1$, such that the amount of context decay, and the amount of new input into context, is maximized, whereas repeated items will likely be activated to some extent in context, resulting in a smaller r value, and therefore less context decay and less input into context.

The input to context in our TCM variant is simply the two presented items, normalized to unit length:

$$\boldsymbol{t^{IN}} = |\boldsymbol{f_A} + \boldsymbol{f_B}|.$$

In previous versions of TCM, $\boldsymbol{t^{IN}}$ has included past context states reinstated from the items (i.e., $\boldsymbol{t'_A}$ and $\boldsymbol{t'_B}$), scaled by a parameter, but for simplicity we limit the new input to the presented items themselves.

Computational model-fitting procedures

We fit the model with DEMC, employing 100 independent chains, each of which included 2500 samples. The first 500 of these samples were used as a burn-in period, and the last 500 iterations of all chains were included in analyses of the posterior distributions. We fit the hierarchical model with Gibbs sampling to update the hyper-priors and differential evolution Markov chain Monte Carlo to perform inference on the subject level (Turner & Van Zandt, 2014).

An important aspect of Bayesian model-fitting approaches is the choice of prior distributions, which add some degree of constraint to the values the parameter can take on. We applied a hierarchical model-fitting procedure, in which each parameter, except for t_0, was fit hierarchically. This allowed us to estimate parameter values for each age group while properly accounting for variability between participants. Specifically, the prior distributions for each participant i were controlled by hyper-parameters specific to each age group j.

$$\begin{aligned}
logit(\rho_{i,j}) &\sim \mathcal{N}(\rho_{\mu_j}, \rho_{\sigma_j})\\
log(\lambda_{i,j}) &\sim \mathcal{N}(\lambda_{\mu_j}, \lambda_{\sigma_j})\\
log(\alpha_{i,j}) &\sim \mathcal{N}(\alpha_{\mu_j}, \alpha_{\sigma_j})\\
log(\kappa_{i,j}) &\sim \mathcal{N}(\kappa_{\mu_j}, \kappa_{\sigma_j})\\
log(\gamma_{i,j}) &\sim \mathcal{N}(\gamma_{\mu_j}, \gamma_{\sigma_j})\\
log(\tau_{i,j}) &\sim \mathcal{N}(\tau_{\mu_j}, \tau_{\sigma_j})\\
log(\nu_{i,j}) &\sim \mathcal{N}(\nu_{\mu_j}, \nu_{\sigma_j})\\
log(a_{i,j}) &\sim \mathcal{N}(a_{\mu_j}, a_{\sigma_j})\\
logit(w_{i,j}) &\sim \mathcal{N}(w_{\mu_j}, w_{\sigma_j})\\
logit(\frac{t_{0_{i,j}}}{min_{RTi}}) &\sim \mathcal{N}(\mu = 0, \sigma = 1.4)
\end{aligned}$$

The mean and standard deviation parameters constraining these parameters were in turn constrained by the following hyper-priors:

$$\begin{aligned}
\rho_{\mu_j} &\sim \mathcal{N}(0, 1)\\
\rho_{\sigma_j} &\sim InvGamma(1, 1)\\
\lambda_{\mu_j} &\sim \mathcal{N}(1, 1)\\
\lambda_{\sigma_j} &\sim InvGamma(1, 1)\\
\alpha_{\mu_j} &\sim \mathcal{N}(1, 1)\\
\alpha_{\sigma_j} &\sim InvGamma(1, 1)\\
\kappa_{\mu_j} &\sim \mathcal{N}(1, 1)\\
\kappa_{\sigma_j} &\sim InvGamma(1, 1))\\
\gamma_{\mu_j} &\sim \mathcal{N}(1, 1)\\
\gamma_{\sigma_j} &\sim InvGamma(1, 1)\\
\tau_{\mu_j} &\sim \mathcal{N}(1, 1)\\
\tau_{\sigma_j} &\sim InvGamma(1, 1)\\
\nu_{\mu_j} &\sim \mathcal{N}(1, 1)\\
\nu_{\sigma_j} &\sim InvGamma(1, 1)\\
a_{\mu_j} &\sim \mathcal{N}(1, 1)\\
a_{\sigma_j} &\sim InvGamma(1, 1)\\
w_{\mu_j} &\sim \mathcal{N}(0, 1)\\
w_{\sigma_j} &\sim InvGamma(1, 1)
\end{aligned}$$

Joint posterior distributions of hyper-parameters

Given that the model contains a relatively high number of free parameters (10), parameter identifiability may be a concern if there could be multiple sets of parameter values that generate the same behavior (Farrell & Lewandowsky, 2018). However, the number of degrees of freedom were much higher than the number of parameters due to fitting trial-level choices and RTs. In addition, we found that the posterior distributions of parameter values were highly constrained relative to the priors, without strong correlations between parameters that could indicate a problem with identifiability (the relationships between each combination of age group-level hyperparameters are shown in Figure S3). Therefore, we have every reason to believe that our model parameters are identifiable.

Fit of alternative models

The data simulated from the alternative models with κ set to zero or 1 are presented along with the observed data in Figure S4.

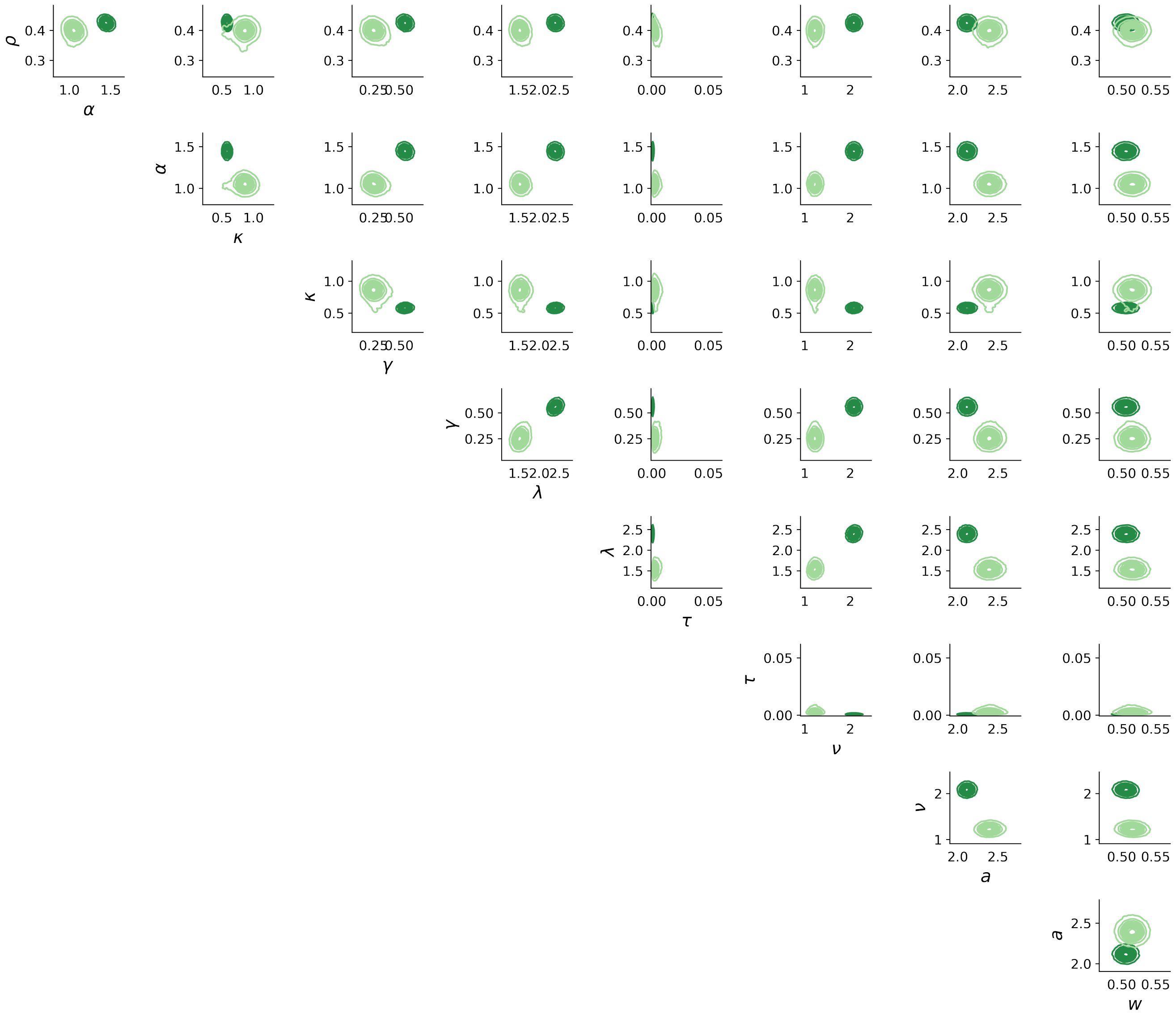

Figure S3. *The joint posterior distributions of mean hyper-parameter distributions.* Each plot shows the density of pairwise combinations of hyper-parameter values for young adults (dark green) and older adults (light green).

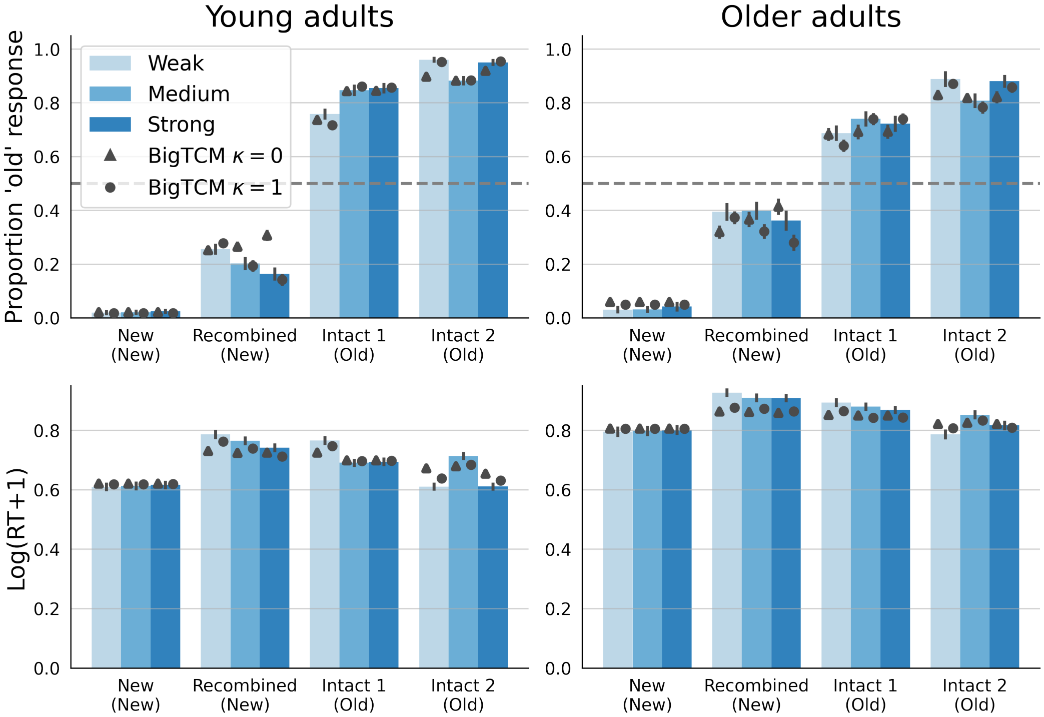

Figure S4. Predictions of the alternative models. Mean observed CAR task performance values are plotted by bars, and mean alternative model-predicted values are presented as triangles and dots. Error bars represent standard errors.