Supplemental Material

Kevin P. Darby
Department of Psychology, Florida Atlantic University

Jessica N. Gettleman, Chad S. Dodson, Per B. Sederberg
Department of Psychology, University of Virginia

Conventional statistical analyses

To assess choices, RTs, and confidence values, we applied a series of analyses of variance (ANOVAs) and t-tests. In post-hoc analyses with multiple comparisons we adjusted the p value with the Benjamini–Hochberg procedure, notated as p_{BH}, which decreases the false discovery rate (Benjamini & Hochberg, 1995). For all assessments, we first compared young and older adults' performance within each experiment separately, followed by a comparison of performance by young and older adults who completed the same procedure across experiments (i.e., the O_1 and Y_1 groups). All of these conventional statistical analyses were performed on data first averaged within each subject.

Item recognition

Figure 2 of the main manuscript presents summary metrics of item recognition and source memory performance, calibration between source memory performance and confidence, and high-confidence source errors, for all groups. The first metric we assessed was the correct rejection rate, or the proportion of trials in which participants accurately responded that a novel statement was "new." An independent samples t-test revealed higher correct rejection rates in older adults than in young adults in Experiment 1 (Y_d v O_1), $t(34) = 4.56, p < .001$, Hedge's $g = 1.49$, suggesting that a 24-hour delay hampered young adults' ability to detect novel statements. In Experiment 2, there was not a significant difference in correct rejection rates between age groups (Y_1 v O_2 v O_3), according to a one-way analysis of variance (ANOVA): $F(2, 51) = 3.03, p = .057, \eta_p^2 = .11$. Similarly, there was not a significant difference between young and older adults who completed the same procedure (Y_1 v O_1), $t(34) = 1.69, p = .10, g = 0.55$.

We also examined hit rates, or the proportion of trials in which participants responded that a studied statement had been associated with either source, regardless of source accuracy. The difference between age groups did not reach significance in Experiment 1 (Y_d v O_1), $t(34) = 2.02, p = .051, g = 0.66$, although a one-way ANOVA revealed a significant difference between groups in Experiment 2 (Y_1 v O_2 v O_3), $F(2, 51) = 26.92, p < .001, \eta_p^2 = .51$. Follow-up tests indicated significant differences between all groups in Experiment 2, such that hit rates were lower in young adults with one stimulus presentation (Y_1) than older adults were either two stimulus presentations (O_2), $t(34) = -4.26, p_{BH} < .001, g = -1.39$, or three presentations (O_3), $t(34) = -6.39, p_{BH} < .001, g = -2.08$. In addition, hit rates were higher in older adults with three presentations (O_3) than with two (O_2), $t(34) = 3.52, p_{BH} = .001, g = 1.15$. These results suggest that additional presentations helped older adults to better recognize the studied items. Indeed, when both young and older adults had one presentation (Y_1 v O_1), there was no difference between groups, $t(34) = 0.40, p = 0.69, g = 0.13$.

Source memory

As explained in the main text, source memory was quantified with paired-source conditional source-identification measure (PSCSIM) scores. These scores did not differ between young and older adults in Experiment 1 according to an independent samples t-test ($Y_d v O_1$): $t(34) = 0.60, p = .56, g = 0.19$. Similarly, a one-way ANOVA revealed that there was no effect of group in Experiment 2 (Y_1 v O_2 v O_3): $F(2, 51) = 0.69, p = .51, \eta_p^2 = .03$. Therefore, the experimental manipulations were sufficient to induce comparable source memory performance between young and older adults in both experiments. However, as expected, when young and older adults completed the same procedure, older adults exhibited a source memory deficit ($Y_1 v O_1$): $t(34) = -4.18, p < .001, g = -1.36$.

RTs for correct rejections and accurate and inaccurate source responses

Figure 3 of the main text presents mean RTs and confidence levels for correct rejections, as well as accurate and inaccurate source responses. For each participant we calculated the mean of log-transformed RTs for correct rejections, accurate source responses, and inaccurate source responses. We first assessed RTs in Experiment 1 with a mixed ANOVA with a 2 (Group: Y_d v O_1) by 3 (Response: correct rejection v correct source v incorrect source) design. Group was treated as a between-subject factor and Response as a within-subject factor. There were main effects of both Group, $F(1, 34) = 6.46, p = .016, \eta_p^2 = .16$, and Response, $F(2, 68) = 43.28, p < .001, \eta_p^2 = .56$, as well as an interaction between these factors, $F(2, 68) = 12.72, p < .001, \eta_p^2 = .27$. To better understand the interaction, we performed post-hoc tests on differences between RTs across response types within

each group. For young adults, RTs were slower for incorrect source responses compared to both correct source responses, $t(17) = 2.55, p_{BH} = .025, g = 0.63$, and correct rejections, $t(17) = 2.74, p_{BH} = .021, g = 0.93$. However, there was no difference in RTs between correct source and correct rejection trials, $t(17) = 0.78, p_{BH} = .45, g = 0.20$. Similar to young adults, older adults were slower for incorrect source responses compared to correct source responses, $t(17) = 4.18, p_{BH} = .001, g = 0.40$, as well as correct rejections, $t(17) = 9.04, p_{BH} < .001, g = 1.53$. Unlike young adults, however, older adults were slower for correct source responses compared to correct rejections, $t(17) = 8.25, p_{BH} < .001, g = 1.10$.

We next performed a similar analysis of the data from Experiment 2, and found the same pattern of results. According to a mixed ANOVA, there was a significant main effect of Group (Y_1 v O_2 v O_3), $F(2, 51) = 10.84, p < .001, \eta_p^2 = .30$, a significant effect of Response, $F(2, 102) = 39.51, p < .001, \eta_p^2 = .44$, and an interaction, $F(4, 102) = 2.52, p = .045, \eta_p^2 = .09$. As for Experiment 1, we investigated the interaction by comparing RTs between different response types for each group. For the young adults (Y_1), RTs were slower for incorrect source compared to correct source responses, $t(17) = 4.97, p_{BH} = .001, g = 0.76$, as well as compared to correct rejection responses, $t(17) = 3.32, p_{BH} = .006, g = 0.66$, whereas there was no difference in RTs between correct source responses and correct rejections, $t(17) = 0.75, p_{BH} = .46, g = 0.13$. For the O_2 group of older adults, mean RTs were slower for incorrect source responses compared to both correct source responses, $t(17) = 3.55, p_{BH} = .004, g = 0.79$, and correct rejections, $t(17) = 5.19, p_{BH} = .001, g = 1.42$. Older adults, unlike young adults, were slower for correct source responses compared to correct rejections, $t(17) = 4.38, p_{BH} = .001, g = 0.70$. Similarly, for older adults in the O_3 group, there were differences in RTs between incorrect and correct source responses, $t(17) = 3.14, p_{BH} = .008, g = 0.77$; between incorrect source and correct rejection responses, $t(17) = 4.20, p_{BH} = .001, g = 1.10$; and between correct source and correct rejection decisions, $t(17) = 2.63, p_{BH} = .02, g = 0.40$.

We also compared young and older adults in the Y_1 and O_1 groups who completed the same procedure. A mixed ANOVA revealed main effects of both age, $F(1, 34) = 15.25, p < .001, \eta_p^2 = .31$, and response type, $F(2, 68) = 44.37, p < .001, \eta_p^2 = .57$, as well as an interaction, $F(2, 68) = 13.47, p < .001, \eta_p^2 = .28$. The interaction was primarily due to differences between correct rejection and correct source responses in older adults, but not young adults, as reported above.

Confidence for accurate and inaccurate source responses

To assess potential differences in confidence between these kinds of responses and between age groups in Experiment 1, we conducted a series of mixed ANOVAs with Group (Y_d v O_1) as a between-subject factor and Response Type (correct rejection v correct source v incorrect source) as a within-subject factor. There was a main effect of Group, $F(1, 34) = 8.34, p = .007, \eta_p^2 = .20$, and a main effect of Response Type, $F(2, 68) = 40.46, p < .001, \eta_p^2 = .54$, but these effects were superseded by an interaction, $F(2, 68) = 18.77, p < .001, \eta_p^2 = .36$. If participants are able to monitor their performance effectively, confidence should be higher for correct responses than incorrect responses. In young adults, confidence was indeed higher for correct source responses compared to incorrect source responses according to a paired t-test, $t(17) = 4.81, p_{BH} < .001, g = 0.62$, and for correct rejections compared to incorrect source responses, $t(17) = 2.97, p_{BH} = .01, g = 0.54$. There was no difference in confidence between correct rejections and correct source trials, however: $t(17) = 0.002, p_{BH} = 1.00, g = 0.00$. Older adults, like young adults, were more confident for correct source compared to incorrect source responses, $t(17) = 4.34, p_{BH} = .001, g = 0.38$, and for correct rejections compared to incorrect source responses, $t(17) = 7.00, p_{BH} < .001, g = 1.95$. Unlike young adults, however, older adults were also more confident in correct rejections compared to correct source trials, $t(17) = 6.04, p_{BH} = .001, g = 1.57$.

We next completed a similar mixed ANOVA on data from Experiment 2, which included the Y_1, O_2, and O_3 groups. Similar to Experiment 1, there were main effects of Group, $F(2, 51) = 7.98, p = .001, \eta_p^2 = .24$, and Response Type, $F(2, 102) = 65.67, p < .001, \eta_p^2 = .56$, as well as an interaction between these factors, $F(4, 102) = 2.57, p = .042, \eta_p^2 = .09$. As in Experiment 1, young adults were less confident in source errors compared to both correct source decisions, $t(17) = -7.65, p_{BH} < .001, g = -1.74$, and correct rejections, $t(17) = -3.50, p_{BH} = .004, g = -1.18$, with no difference between correct source and correct rejection responses, $t(17) = -0.30, p_{BH} = .77, g = -0.07$. Older adults in the O_2 were also less confident in source errors than for both correct source responses, $t(17) = -5.39, p_{BH} < .001, g = -0.80$, and correct rejections, $t(17) = -7.23, p_{BH} < .001, g = -1.91$. Unlike young adults, but consistent with results from Experiment 1, older adults in the O_2 group were less confident for correct source compared to correct rejection responses, $t(17) = -5.95, p_{BH} < .001, g = -1.27$. The same pattern was observed for the O_3 group, in which confidence was lower for source error than for source correct responses, $t(17) = -3.27, p_{BH} = .005, g = -0.66$, and for correct rejections, $t(17) = -5.65, p_{BH} < .001, g = -1.59$, and were also less confident in correct source compared to correct rejection decisions, $t(17) = -3.93, p_{BH} = .002, g = -1.05$.

Finally, we compared young and older adults who completed the same procedure by including the Y_1 and O_1 groups in a mixed ANOVA. There was not a main effect of Group for this comparison, $F(1, 34) = 3.12, p = .086, \eta_p^2 = .08$, suggesting similar levels of confidence overall between young and

older adults. However, there was a main effect of Response, $F(2, 68) = 42.52, p < .001, \eta_p^2 = .56$, and an interaction between the factors, $F(2, 68) = 9.43, p < .001, \eta_p^2 = .22$. The interaction was primarily due to differences in confidence between correct source and correct rejection decisions in older adults, but not in young adults, as presented above.

Proportion of high confidence source errors

A key prediction of the misrecollection account is that older adults should exhibit a higher proportion of high confidence responses for source memory errors. To test this, we selected the trials on which each participant chose the incorrect source option for studied statements, and calculated the proportion of those trials on which either a 9 or 10 was selected on the confidence scale (see Figure 2). In Experiment 1, there was not a statistically significant difference between the Y_d and O_1 groups: $F(1, 34) = 3.52, p = .07, \eta_p^2 = .09$. The difference was significant overall in Experiment 2, $F(2, 51) = 4.42, \eta_p^2 = .15$, arising from differences between young adults in the Y_1 group and older adults in both the O_2 group, $t(17) = 3.13, p_{BH} = .01, g = 1.02$, and the O_3 group, $t(17) = 2.79, p_{BH} = .01, g = 0.99$. There was also a significant difference between young adults and older adults who performed the same procedure (i.e., Y_1 and O_1): $F(1, 34) = 8.65, p = .006, \eta_p^2 = .20$. Overall, then, the results replicate findings of more high-confidence errors in older adults.

Calibration between source memory accuracy and confidence

Figure 2 also shows the calibration between source performance and confidence in each group. This is calculated according to Equation 1 of the main text. Higher values of this metric indicate greater error in the calibration.

An independent samples t-test revealed no effect of age on calibration in Experiment 1 (Y_d v O_1): $t(34) = 0.89, p = .38, g = 0.29$. However, there was a significant effect of group in Experiment 2 (Y_1 v O_2 v O_3), $F(2, 51) = 3.97, p = .02, \eta_p^2 = .13$, which was driven by greater calibration error in the O_2 group compared to Y_1: $t(34) = 3.17, p_{BH} = .01, g = 1.03$. The other pairwise comparisons did not reach significance after correcting for multiple comparisons ($p_{BH} > .19$). When young and older adults performed the same procedure (Y_1 v O_1), older adults exhibited greater calibration error: $Y_1 v O_1, t(34) = 2.76, p = .01, g = 0.90$.

Computational model

The model is described in detail in the main manuscript. Below, we provide visualizations of differences between the four sigmoid model variants described in the main text, and specify the prior distributions we applied to fit the models. We also provide figures showing predictions of the standard RBOE model.

Visualizing the sigmoid model variants

As described in the main text, we compared four model variants that used variations of a sigmoid function for the mapping between accumulated evidence – specifically the distance between evidence and the corresponding decision threshold – and subjective confidence.

Figure S1 visualizes these four model variants. The plot on the left shows how evidence for three different choices accumulated across time for a simulated trial, whereas the other plots show how confidence would change were the evidence values for the two losing choices different in each model variant. In each plot, the green curve shows how confidence would change given different values of the green accumulator, constrained by the final value of the orange accumulator as is shown in the left-side figure. In the same way, the orange line shows how confidence would change with different values of that accumulator given the actual final value of the green accumulator. The orange and green dots represent the confidence value given the level of evidence that was observed for each accumulator.

Consider the separate sigmoid model in Figure S1. Because the green accumulator was relatively close to the threshold when the blue accumulator reached the threshold, high confidence would not be possible despite changing the distance of the orange accumulator. Changing the distance of the green accumulator, however, could have a larger effect on confidence, since the distance of the orange accumulator is relatively high. For the summed sigmoid model, in contrast, high confidence is possible regardless of the evidence for the green accumulator since the distances are summed across accumulators. For the minimum and maximum sigmoid models, only one distance has an impact on confidence – either the smallest distance (in the minimum sigmoid model) or the largest distance (in the maximum sigmoid model).

Parameter priors

Because the model parameters were fit hierarchically (except for t_0), the priors for participant-level parameters came from hyperpriors that were specific to each dataset (i.e., Y_d, O_1, Y_1, O_2, and O_3). The participant-level parameters (including parameters fit only in rejected model variants, as described in the main manuscript) were fit with normal distribution priors as follows:

$$log(\rho_{n_{j,k}}) \sim \mathcal{N}(\mu_{\rho_{n_j}}, \sigma_{\rho_{n_j}})$$

$$log(\rho_{f_{j,k}}) \sim \mathcal{N}(\mu_{\rho_{i_j}}, \sigma_{\rho_{i_j}})$$

$$log(\rho_{s_{j,k}}) \sim \mathcal{N}(\mu_{\rho_{s_j}}, \sigma_{\rho_{s_j}})$$

$$logit(\phi_{j,k}) \sim \mathcal{N}(\mu_{\phi_j}, \sigma_{\phi_j})$$

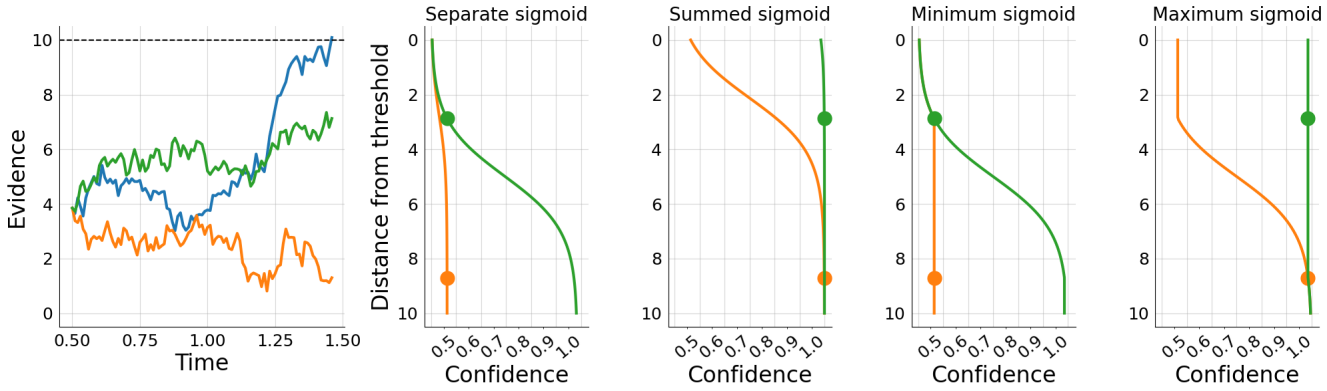

Figure S1. Confidence calculations of the separate, summed, minimum, and maximum sigmoid models. The plot on the left depicts evidence accumulation in a 3-choice task. The lines on the four plots on the right depict how confidence would change according to each model when varying the distance of one accumulator and keeping the other distance constant. For example, the green lines show confidence would change given any level of evidence for the green accumulator, constrained by the observed level of evidence for the orange accumulator (see the text for details). The green and orange dots represent confidence given the distances observed in the plot on the left.

$$log(\kappa_{j,k}) \sim \mathcal{N}(\mu_{\kappa_j}, \sigma_{\kappa_j})$$
$$log(\beta_{j,k}) \sim \mathcal{N}(\mu_{\beta_j}, \sigma_{\beta_j})$$
$$log(\tau_{j,k}) \sim \mathcal{N}(\mu_{\tau_j}, \sigma_{\tau_j})$$
$$log(\delta_{j,k}) \sim \mathcal{N}(\mu_{\delta_j}, \sigma_{\delta_j})$$
$$log(\alpha_{new_{j,k}}) \sim \mathcal{N}(\mu_{\alpha_{new_j}}, \sigma_{\alpha_{new_j}})$$
$$log(\alpha_{prop_{j,k}}) \sim \mathcal{N}(\mu_{\alpha_{prop_j}}, \sigma_{\alpha_{prop_j}})$$
$$logit(\frac{t_{0_{j,k}}}{min_{RT_{j,k}}}) \sim \mathcal{N}(0, 1.4),$$

where j is the dataset and k is the participant. These subject-level priors (excluding the prior for t_0) were governed by group-level hyper-parameters for the means and standard deviations. The hyper-parameters were controlled by the following hyperpriors:

$$\mu_{\rho_{n_j}} \sim \mathcal{N}(-0.5, 2)$$
$$\sigma_{\rho_{n_j}} \sim InvGamma(2, 1)$$
$$\mu_{\rho_{f_j}} \sim \mathcal{N}(-0.5, 2)$$
$$\sigma_{\rho_{f_j}} \sim InvGamma(2, 1)$$
$$\mu_{\rho_{s_j}} \sim \mathcal{N}(-0.5, 2)$$
$$\sigma_{\rho_{s_j}} \sim InvGamma(2, 1)$$
$$\mu_{\phi_j} \sim \mathcal{N}(0, 2)$$
$$\sigma_{\phi_j} \sim InvGamma(2, 1)$$
$$\mu_{\kappa_j} \sim \mathcal{N}(-0.5, 2)$$
$$\sigma_{\kappa_j} \sim InvGamma(2, 1)$$
$$\mu_{\beta_j} \sim \mathcal{N}(-0.5, 2)$$
$$\sigma_{\beta_j} \sim InvGamma(2, 1)$$
$$\mu_{\tau_j} \sim \mathcal{N}(1, 2)$$
$$\sigma_{\tau_j} \sim InvGamma(2, 1)$$
$$\mu_{\delta_j} \sim \mathcal{N}(2, 2)$$
$$\sigma_{\delta_j} \sim InvGamma(2, 1)$$
$$\mu_{\alpha_{new_j}} \sim \mathcal{N}(2, 2)$$
$$\sigma_{\alpha_{new_j}} \sim InvGamma(2, 1)$$
$$\mu_{\alpha_{prop_j}} \sim \mathcal{N}(1, 2)$$
$$\sigma_{\alpha_{prop_j}} \sim InvGamma(2, 1)$$

Comparing the fit of the summed sigmoid and baseline RBOE models

As we demonstrate in the main manuscript, the summed sigmoid model was preferred over all other variants. The relative balance of evidence (RBOE) models provided particularly poor fits of the data, and the curious reader may desire to better understand why this method of calculating confidence was less able to account for the data. Figure 8 of the main text shows the mean proportion of every level of confidence for each response type and each dataset, for both the summed sigmoid and standard RBOE models, and it is clear that the summed sigmoid model was better able to match the proportions of confidence values than the RBOE model. Figures S2 and S3 depict the predictions of the standard RBOE model in the same way that Figures 2 and 3 of the main manuscript depict the predictions of the summed sigmoid model. The RBOE model only differed from the summed sigmoid model in how confidence was calculated, and it is clear from Figure S3 that confidence judgments were

less flexible in this model, resulting in less distinct differences in confidence between response types. For example, RBOE was less able to account for reduced confidence for incorrect source responses compared to correct source responses.

Interestingly, the RBOE model did not only miss the mark when it came to confidence judgments, but was less able to fit some aspects of the choice and RT data as well, such as under-predicting correct rejections in some cases (Figure S2) and often predicting substantially too-fast RTs (Figure S3). Why would the RBOE model make different predictions compared to the sigmoid model in performance areas other than confidence? The answer is that other model parameters needed to adjust their values to try to provide as much flexibility to confidence judgments as possible, such as by adjusting relative threshold and drift rate values, since the mapping between evidence and confidence has no flexibility in the RBOE framework.

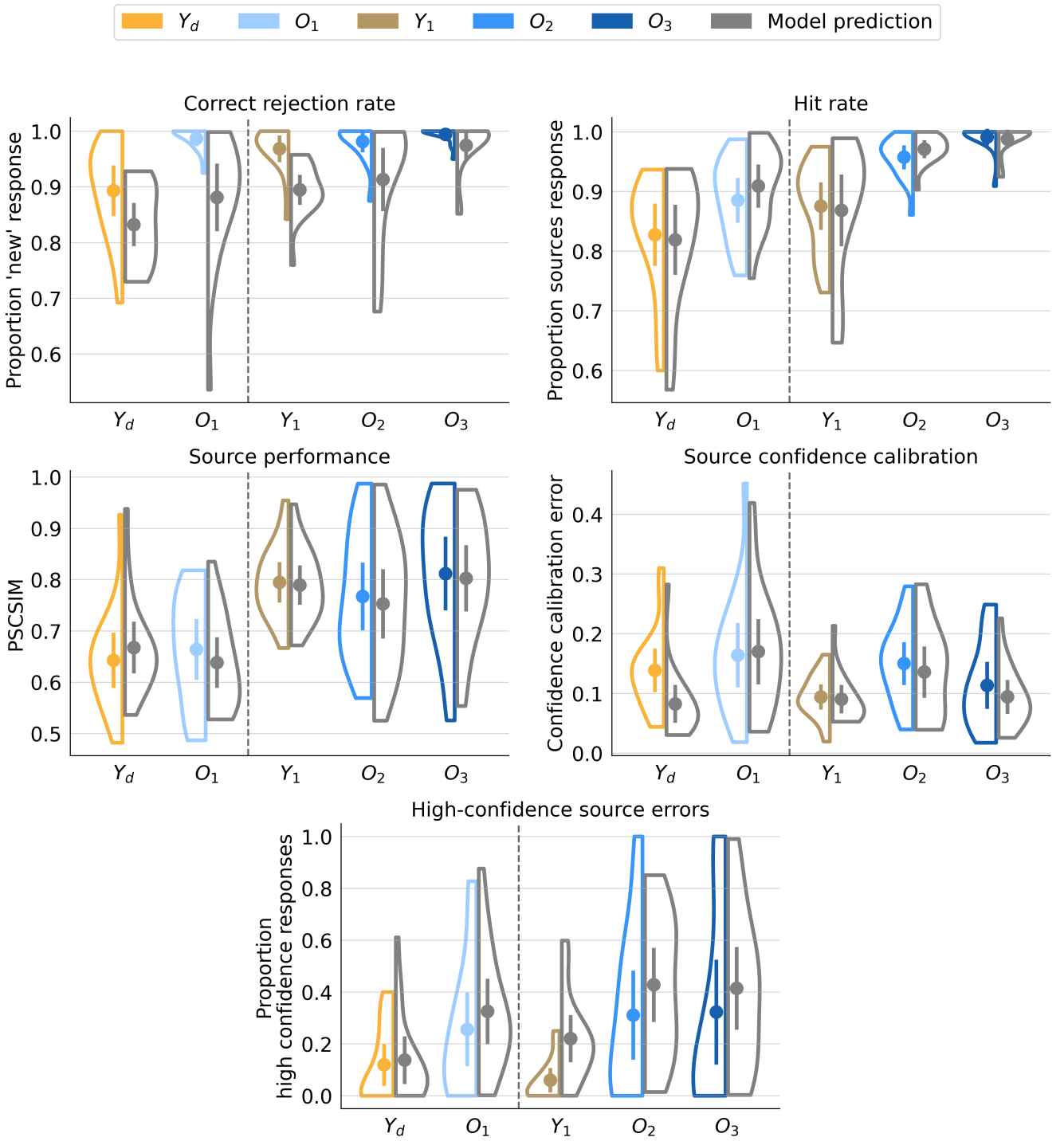

Figure S2. Comparison of observed and standard RBOE model-predicted response accuracy and accuracy-confidence calibration results. Each split violin shows the distribution of observed performance on the left, along with mean level of performance, for each group. The gray distribution on the right side of each split violin shows the performance simulated by the winning computational model for each participant, along with the mean across participants. The vertical dashed line on each plot separates the datasets of Experiments 1 and 2. The error bars indicate the 95% confidence interval for each distribution.

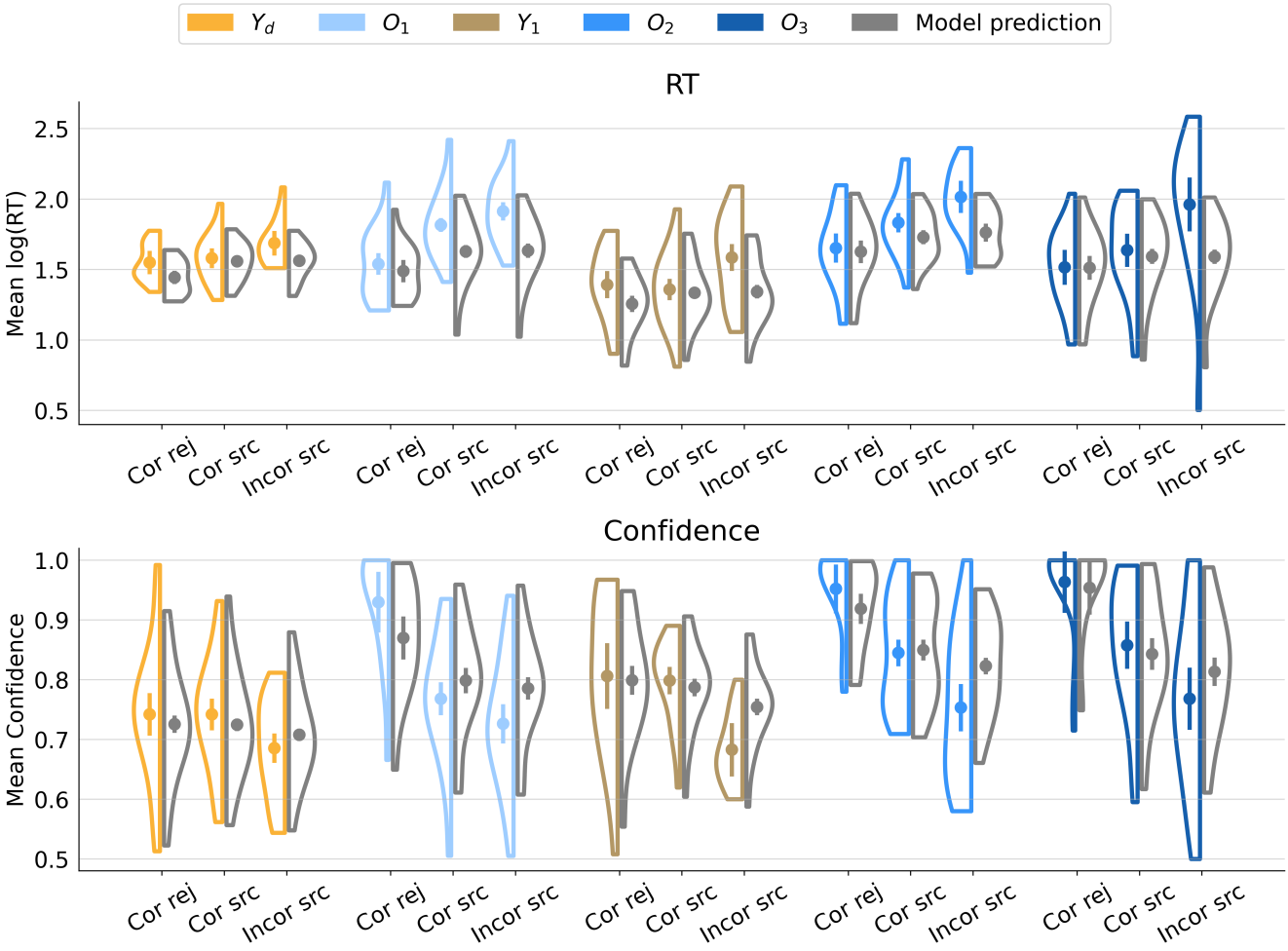

Figure S3. Comparison of observed and standard RBOE model-predicted mean response times (top) and confidence judgments (bottom) for correct rejections, correct source responses, and incorrect source responses. Each split violin shows the distribution of observed performance on the left, along with mean level of performance, for each group. The gray distribution on the right side of each split violin shows the performance simulated by the winning computational model for each participant, along with the mean across participants. The error bars indicate the 95% confidence interval for each distribution, corrected for within-subject comparisons within each group. Cor rej = correct rejections; Cor src = correct source; Incor src = incorrect source.