Novelty rejection in episodic memory - Supplemental Materials

Adam F. Osth, Aspen Zhou, Simon Lilburn, & Daniel R. Little

University of Melbourne

Address correspondence to:

Adam Osth (E-mail: adamosth@gmail.com)

Novelty rejection in episodic memory - Supplemental Materials

In these Supplemental Materials, we report on additional analyses, models, and an additional experiment.

A: Experiment 3B: Equating Distance between Standard and Extralist Feature Lures

In Experiment 3 in the main article, 1/distance was equated between the standard and extralist feature lures. We additionally ran an experiment that equated the distance between the probe types, which we report here. Data from this experiment can be found on our OSF page (`https://osf.io/b2zyk/`.

Method

Participants. Participants were 18 members of the University of Melbourne community, with normal or correct-to-normal vision, who participated in three one-hour sessions. Participants were renumerated at a rate $15 Australian dollars per session. Human testing was approved by the Melbourne Human Research Ethics Committee (Approval number: 1034866).

Materials and Procedure

Both the stimuli and procedure were identical to Experiment 3 in the main text. Study sets and test items were constructed in the same manner as Experiment 3, except that we constrained lures and extralist feature lures have equivalent values for 1 over the summed distance to each study set item. Distance was also computed using the city-block metric.

Results

Results can be seen in Figure 1. Like in Experiments 1 and 2, an effect was found for lure distance, $BF_{10} = 3.093 \times 10^{1}4$, with distance-1 lures having much higher FAR

($M = .546, SEM = .023$) than lures at distance-2 ($M = .343, SEM = .026$). An effect of lure type was found, $BF_{10} = 2.29e + 15$, with ELF probes showing much *higher* FAR ($M = .548, SEM = .023$) than standard lures ($M = .341, SEM = .029$). The dimension that was fixed in the study set had no effect on the FAR, $BF_{10} = .055$, with roughly equivalent FAR for each fixed dimension (saturation $M = .456, SEM = .027$, height $M = .439, SEM = .024$, bar position $M = .438, SEM = .023$).

An interaction between distance and the dimension that held the fixed dimension were both found, as the model with this interaction and all three main effects was preferred, $BF_M = 9.29$. This interaction is not of theoretical interest and was not analyzed further.

B: Parallel and Hybrid Model Fits to Experiment 3's Data

Parallel Models

The main text depicts the fits of a parallel model with an exhaustive rule for "yes" decisions and a self-terminating rule for "no" decisions. While this was the best performing parallel model, we additionally implemented models with a.) two required "yes" thresholds and a self-terminating rule for "no" decisions, b.) an exhaustive rule for "yes" decisions and two required "no" thresholds, and c.) exhaustive decision rules for both "yes" and "no" decisions. Fits of each of these models to Experiment 2's data can be found in Figure 2.

One can see that these model predictions bear a qualitative resemblance to the self-terminating model in the main text. In particular, each model predicts an extralist feature effect that is larger for distance-2 lures. Nonetheless, the models differ in their quantitative predictions. For instance, the exhaustive model appears to best capture the extralist feature effect for the distance-1 lures, whereas the other two parallel models predict a very small effect for such lures. However, the exhaustive model appears to perform more poorly on the RTs than the other two models. Specifically, it predicts a very large extralist feature effect in the RTs - RTs are much faster for ELF lures than standard lures, whereas the data show much smaller differences.

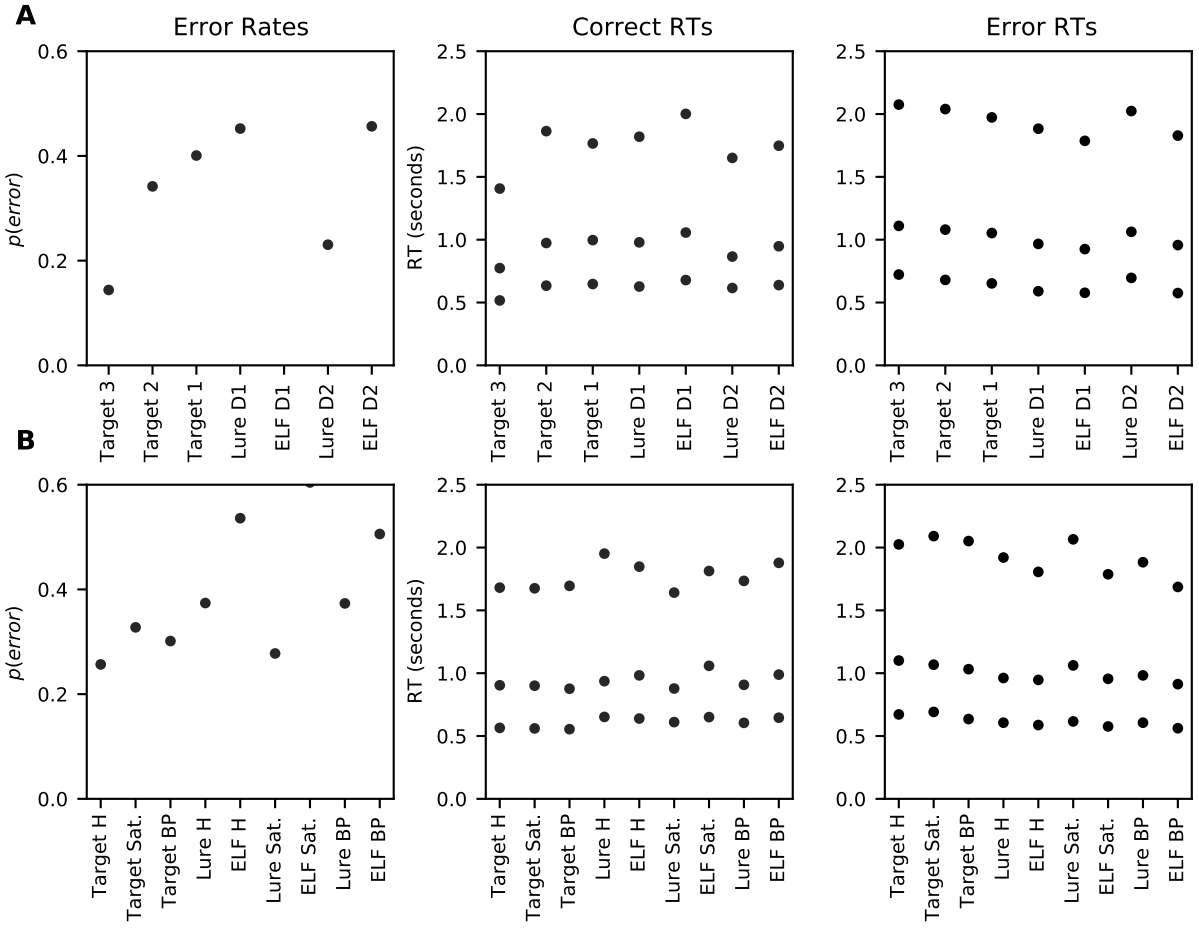

Figure 1. Group-averaged error rates (left panel) and correct and error RT distributions (middle and right panels, respectively) for the data from Experiment 3b with separable-dimension stimuli. The top row (A) shows the results for targets from each serial position and lures from each level of distance. The bottom row shows the results for targets and lures for each fixed dimension collapsed across lure distance and target serial position. RT distributions are summarized using the .1, .5, and .9 quantiles. Notes: D = distance, Sat. = saturation, H = height, BP = bar position.

Hybrid Coactive-Parallel Models

The main text reports on the results of the winning hybrid coactive-parallel model, which is the model where two "new" accumulators hitting the threshold are required to

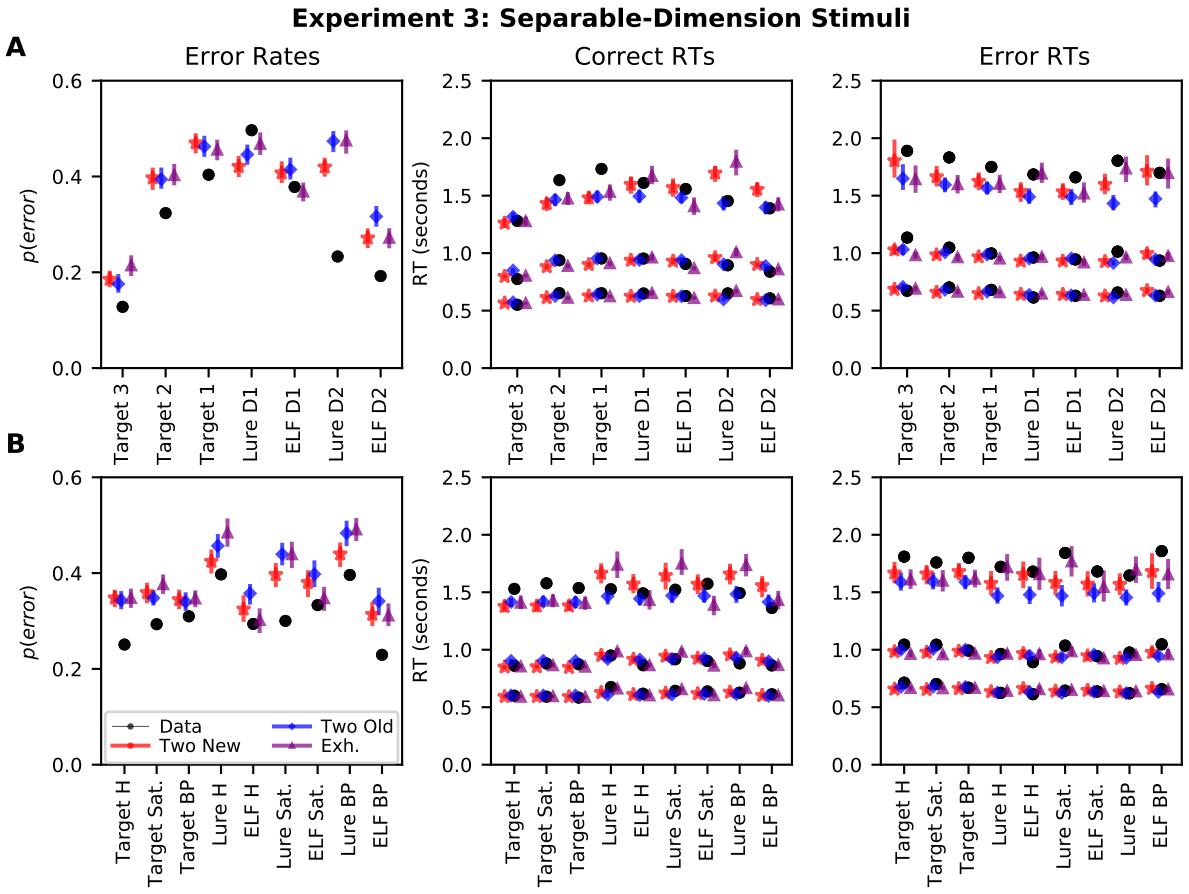

Figure 2. Group-averaged error rates (left panel) and correct and error RT distributions (middle and right panels, respectively) for the data from Experiment 3 with separable-dimension stimuli. RT distributions are summarized using the .1, .5, and .9 quantiles. The top row (A) shows targets separated by serial position and lure results separated by distance. The bottom row (B) shows the results separately for each fixed dimension in the probe stimulus. Model predictions are group-averaged posterior predictives from the parallel global similarity models with different decision rules, including the model where two "new" responses are required (red), the model where two "old" responses are required (blue), and the exhaustive model (purple). Error bars depict the 95% highest density interval (HDI). Note: D1 = distance 1, D2 = distance 2, H = height, Sat. = saturation, BP = bar position.

make a "new" response. However, we additionally implemented two other models that did not fare as well in the model selection procedure, namely a model that uses a self-terminating decision rule for "new" responses along with a model that uses an exhaustive decision rule.

Fits of each of these models to the data from Experiment 2 can be seen in Figure 3. The self-terminating model successfully accounts for the small extralist feature effect for the distance-2 lures. However, unfortunately it predicts a much smaller effect for the distance-1 lures, whereas the data show the opposite pattern. Somewhat interestingly, the exhaustive model predicts a much larger extralist feature effect. However, similar to the pure parallel exhaustive model described above, this comes at the cost of predicting an extralist feature effect on RTs that is much larger than is seen in the data.

C: Parallel and Hybrid Model Fits to Experiment 4's Data

Due to the failures of the parallel models in capturing the separable-dimension stimuli in Experiment 3, Experiment 4 places considerably more focus on the predictions of the core EB-LBA model and the diagnostic attention variant, the latter of which succeeded in accounting for many aspects of the data. Here, we depict the fits of the parallel and hybrid models that were omitted from the main text.

Parallel Models

As with Experiment 3, four parallel EB-LBA models were applied to the data with various decision rules. These include a model where a.) an exhaustive rule for "old" responses and a self-terminating rule is used for "new" responses, b.) an exhaustive rule for "old" responses and two "new" accumulators have to hit the threshold to produce a "new" response, c.) two "old" accumulators have to hit the threshold to produce an "old" response and a self-terminating rule is used for "new" responses, and d.) an exhaustive rule is used for both "old" and "new" responses.

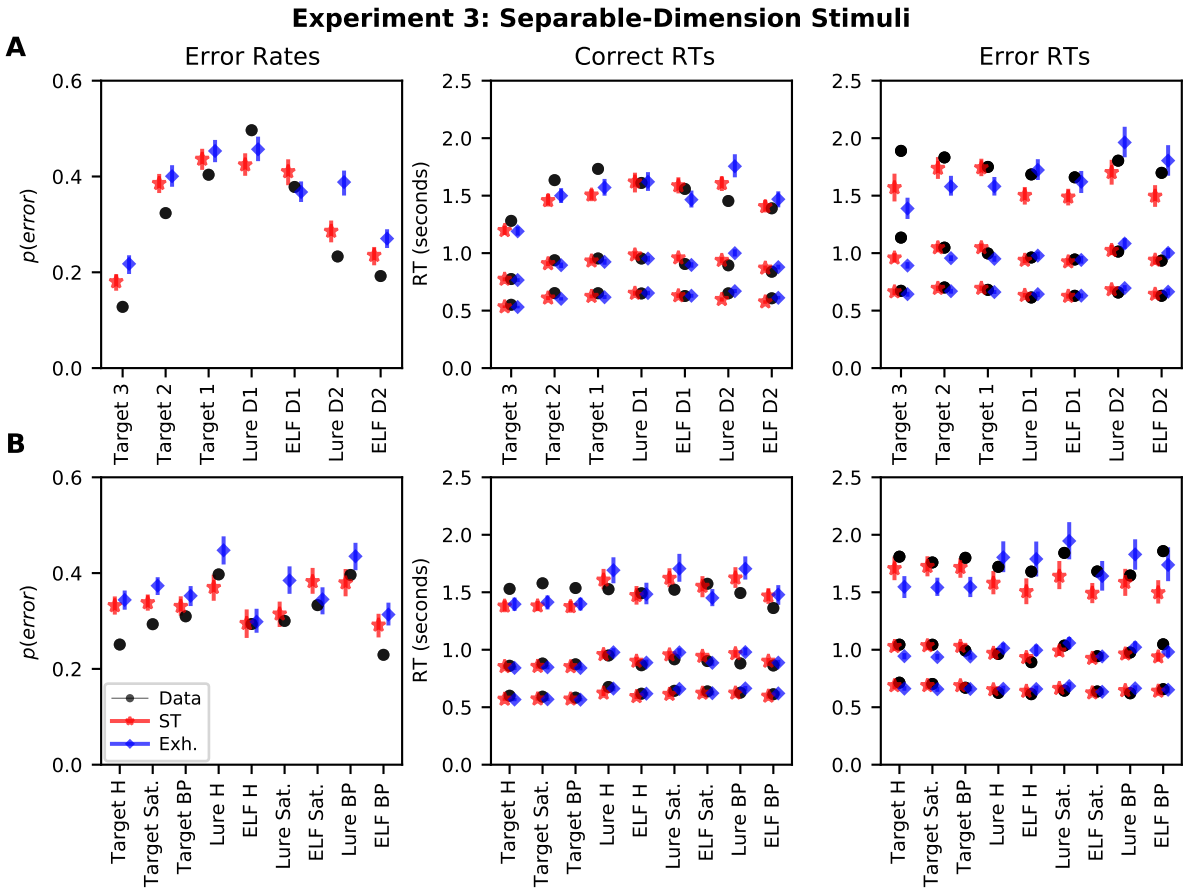

Figure 3. Group-averaged error rates (left panel) and correct and error RT distributions (middle and right panels, respectively) for the data from Experiment 3 with separable-dimension stimuli. RT distributions are summarized using the .1, .5, and .9 quantiles. The top row (A) shows targets separated by serial position and lure results separated by distance. The bottom row (B) shows the results separately for each fixed dimension in the probe stimulus. Model predictions are group-averaged posterior predictives from the hybrid coactive-parallel EB-LBA models with different decision rules, including the self-terminating model (red) and the exhaustive model (blue). Error bars depict the 95% highest density interval (HDI). Note: D1 = distance 1, D2 = distance 2, H = height, Sat. = saturation, BP = bar position.

Fits of a.) and b.) can be seen in Figure 4. Surprisingly, the self-terminating model captured the error rates very well. The model captured all of the qualitative trends in the data, including a near-equivalence in FAR to standard lure and same-dimension ELF (S-ELF) lures at distance-1, albeit with a slightly elevated FAR to the S-ELF lures. In addition, the model captured the reduced FAR to S-ELF lures at distance-2, and the reduced FAR to ELF lures at both distance levels. The model also gets quite close to the data in each of these comparisons. The only limitation in the error rate predictions is that the model underpredicts the error rate for targets at each serial position. Aside from that limitation, the model also falls short of capturing the RTs for targets at serial position 2 and 3, and additionally underpredicts the error RTs for the lures. The diagnostic attention model (depicted in the main text) does not share any of these shortcomings.

However, despite the impressive success of the self-terminating model, the other parallel models fared quite poorly in their ability to capture the data. The model where two "new" accumulators are required to hit the threshold (depicted in Figure 4) did not capture error rates very well, and noticeably predicted nearly equivalent FAR for all lures at distance-2. Fits of c.) and d.) can be seen in Figure 5, which reveals that these alternative parallel models did not fare much better. Both models considerably overpredicted the FAR to the relatively easy distance-2 lures, and did not fare very well in their ability to capture the RTs. The model where two "old" accumulators are required to hit the threshold generally produced faster RTs than are seen in the data across many different stimulus types.

Hybrid Models

A total of three hybrid coactive-parallel models were fit to the data. In each model, a coactive decision rule was employed for "old" responses while the decision rule for the parallel accumulators for "new" responses was varied. These include a.) a self-terminating rule, b.) two "new" accumulators are required to hit the threshold to produce a "new"

decision", and c.) an exhaustive decision rule.

Fits of each of the models to the data from Experiment 4 can be found in Figure 6. Similar to the parallel models, the self-terminating model performed best in its ability to capture the error rates, and gets close to the data in many comparisons. However, it appears to perform more poorly than the purely parallel self-terminating model seen above – while it shares the underprediction of the hit rate, it also overpredicts the FAR to the distance-2 lures. In addition, it generally predicts error RTs to lures that are faster than seen in the data.

The other two hybrid models fare worse in their ability to capture the data. In particular, they overpredict the FAR to the distance-2 lures to a substantial degree. The model where two "old" accumulators are required to hit the threshold performs even worse than the other two models in its ability to capture the hit rates to targets. In addition, the two models overpredict the RTs for lures. Neither of these models appear to provide a satisfactory account of the data.

D: Fits of the Attention to Unvarying Dimensions Model

In the main text, an alternative to the diagnostic attention model is presented where attention shifts to dimensions within the study set that show little variance in their values across the study list – the attention to unvarying dimensions model. This model critically differs from the diagnostic attention model because the values within the probe do not influence how selective attention is allocated.

Fits of the attention to unvarying dimensions model are presented in Figure 7, where the model's predictions are presented alongside the diagnostic attention model, which happens to be the preferred model. One can see that both models capture all of the qualitative predictions and perform very similarly to each other. The advantage of the diagnostic attention model is admittedly quite subtle – one can see that the model does a slightly better job of capturing the group-averaged error rates. This is likely due to the fact

that the diagnostic attention model shifts attention across the probe types, giving extra attention to the fixed dimension only on ELF trials. The attention to unvarying dimension model instead invariably allocates the most attention to the fixed dimension.

Clearer differences between the models can be seen in the fits to Experiment 4's data, which can be seen in Figure 8. While one can again see that the two models perform very similarly to each other, a more crucial difference concerns the predictions for the same-dimension (S-ELF) trials. Recall from the main text that S-ELF trials critically differ from ELF trials because the extralist feature is not carried by the fixed dimension, but is instead carried by a dimension where the values of the study set items vary from one item to another. An advantage for S-ELF probes was found for distance 2, but not for the distance 1 probes. A clear reason for this discrepancy was given by the diagnostic attention model – the extralist feature within the distance 1 S-ELF trials is within the distribution of study set values, indicating a low level of novelty, such that additional attention is not allocated to that dimension. However, for S-ELF trials at distance 2, the extralist feature is considerably more novel, resulting in additional attention and a rejection advantage for such trials.

The attention to unvarying dimensions model cannot capture the interaction between lure type and lure distance. Figure **??** reveals that the model predicts equal FAR to standard lures and S-ELF lures at both levels of lure distance. While this is the clearest disadvantage for the attention to unvarying dimensions model, the diagnostic attention model also appears to provide a stronger account of the RT distributions.

E: Parallel and Hybrid Model Fits to Experiment 5's Data

Parallel Global Similarity Model

The usage of discrete features again leads to some differences in how the parallel model is implemented. The similarity for a particular aspect (color or shape) k is:

$$s_{ijk} = w_k \tau_k \tag{1}$$

Activation values and drift rates are calculated according to Equations **??** and **??**. Due to only two aspects being present in the stimuli, a total of four accumulators implement decisions (two "old" and two "new" accumulators). The parallel model had the same number of parameters as the core EB-LBA model (11).

The usage of only two aspects in the parallel model limits our consideration of model variants. In our investigation, we considered a.) a model with an exhaustive rule for "old" decisions and a self-terminating rule for "new" decisions, which captures the idea that novelty rejection requires less information than acceptance of targets, and b.) a model with an exhaustive rule for both types of decisions. Table **??** reveals that the parallel model with a self-terminating rule for "new" decisions yields a sizeable WAIC advantage over the exhaustive variant ($\Delta_{WAIC} = 1324$), so our coverage here focuses on that particular model.

Group-averaged posterior predictives of the selected parallel model can be seen in Figure **??**A. Similar to our explorations with continuous-dimension stimuli, the parallel model succeeds in predicting the extralist feature effect – much lower FAR are predicted for 2:0 probes than for 1:1 probes. This success is also shared in its account of the RT distributions, where the model does an impressive job of capturing the significantly slower correct rejections of 1:1 probes. These differences occur for the same reason as with continuous dimension stimuli – while only a single rejection of a feature is required to reject the probe, the once-presented features in 1:1 probes have moderate levels of global similarity, whereas the never-presented feature in 2:0 probes has very low global similarity, leading to fast and frequent "no" decisions for this feature.

Unfortunately, a major failure of the model is that it cannot discriminate targets from 1:1 lure probes. This is due to the fact that in a pure parallel model, decisions are only based on whether there is memory for each stimulus aspect (color or shape) – there is no sensitivity to the conjunction of these features. This problem is alleviated in the

coactive-parallel hybrid model, which is described next. The inability to properly discriminate targets from lures is likely why the parallel model performed poorly relatively to the core EB-LBA model in model selection ($\Delta_{WAIC} = 575$).

Coactive-Parallel Hybrid Architecture. In the coactive-parallel hybrid EB-LBA model, evidence for "old" decisions stems from the global similarity of the entire stimulus (the similarity calculations in Equation **??**), while "new" decisions" stems from rejecting the individual features of the stimulus (Equation 1). Thus, the hybrid model combines the strength of each approach. The fact that "new" decisions are based on the global similarity of the individual features makes it such that unseen features can be quickly and easily rejected, facilitating rejections of lures with extralist features.

Following our previous applications of the hybrid model in Experiments 3 and 4, we allowed the LBA parameters to vary across the two decision architectures (sv, a, and t_0). The hybrid model contained a total of 15 parameters per participant. We pursued two hybrid architectures – one which employed a self-terminating decision rule for "new" responses and another model which used an exhaustive rule. Table **??** reveals that the former model yielded a significant advantage in model selection ($\Delta_{WAIC} = 568$) and thus is the variant of focus in the main article.

Group-averaged posterior predictives of the selected hybrid model can be seen in Figure **??**. While the hybrid model struggled with several aspects of the data with continuous-stimuli in Experiments 3 and 4, it succeeded in capturing several qualitative aspects of the data here. It predicted a substantial extralist feature effect – 2:0 probes exhibited much lower FAR than 1:1 probes, with the model even capturing the much slower rejections to 1:1 lures. The model also predicted a small rejection advantage for 1:0 lures over 2:0 lures that is similar in size to that found in the data. The model succeeded in capturing these benchmarks while successfully discriminating targets from 1:1 lures. These strengths of the model likely are the reason why the model yielded an advantage in model selection over the core EB-LBA model ($\Delta_{WAIC} = 115$).

Nonetheless, there are still some shortcomings in the model's ability to capture the data. Error responses to extralist lures (1:0 and 2:0) are considerably slower than the model predicts. This is likely due to the fact that error responses to lures are "old" responses, which rely on the coactive architecture, which is generally unable to account for the wide variation of RTs in the data (e.g., very fast hits to recent targets and very slow false alarms to extralist lures). Figure **??**B also reveals that the hybrid model predicts only slightly higher FAR to lures when shape carries the extralist dimension. In contrast to the coactive EB-LBA model, attention was less focused on the color aspect ($w^{\mu} = .521$, $95\%HDI = [.461, .580]$).

Why did the hybrid model succeed in capturing so many of the benchmarks in this experiment, whereas it performed relatively poorly with our continuous-dimension stimuli in Experiments 3 and 4? While the reasons are not immediately clear, there are several important differences that should be considered. First, our experiments with continuous-dimension stimuli employed a fairly rich set of constraints, where stimulus difficulty (e.g., lure distance) and the presence or absence of extralist features (e.g., standard lures and ELF lures) had strong independent effects. The difficulty manipulation in the present experiment (2:0 vs 1:0 lures) yielded a relatively small effect, and there was no corresponding difficulty manipulation for lures which lacked extralist features (such as 2:2 or 2:1 lures) in order to be consistent with the Mewhort and Johns designs.

In addition, our continuous-dimension stimuli contained three dimensions in Experiments 3 and 4, yielding a total of three accumulators for "new" decisions in those experiments as opposed to the two accumulators in the present experiment. Because of the stochastic nature of the accumulators, additional accumulators produce more decision noise, which can lead to erroneous false recognition of the stimulus dimensions. Future work may be required to test the specific conditions when a hybrid-coactive parallel architecture succeeds in capturing extralist feature effects.

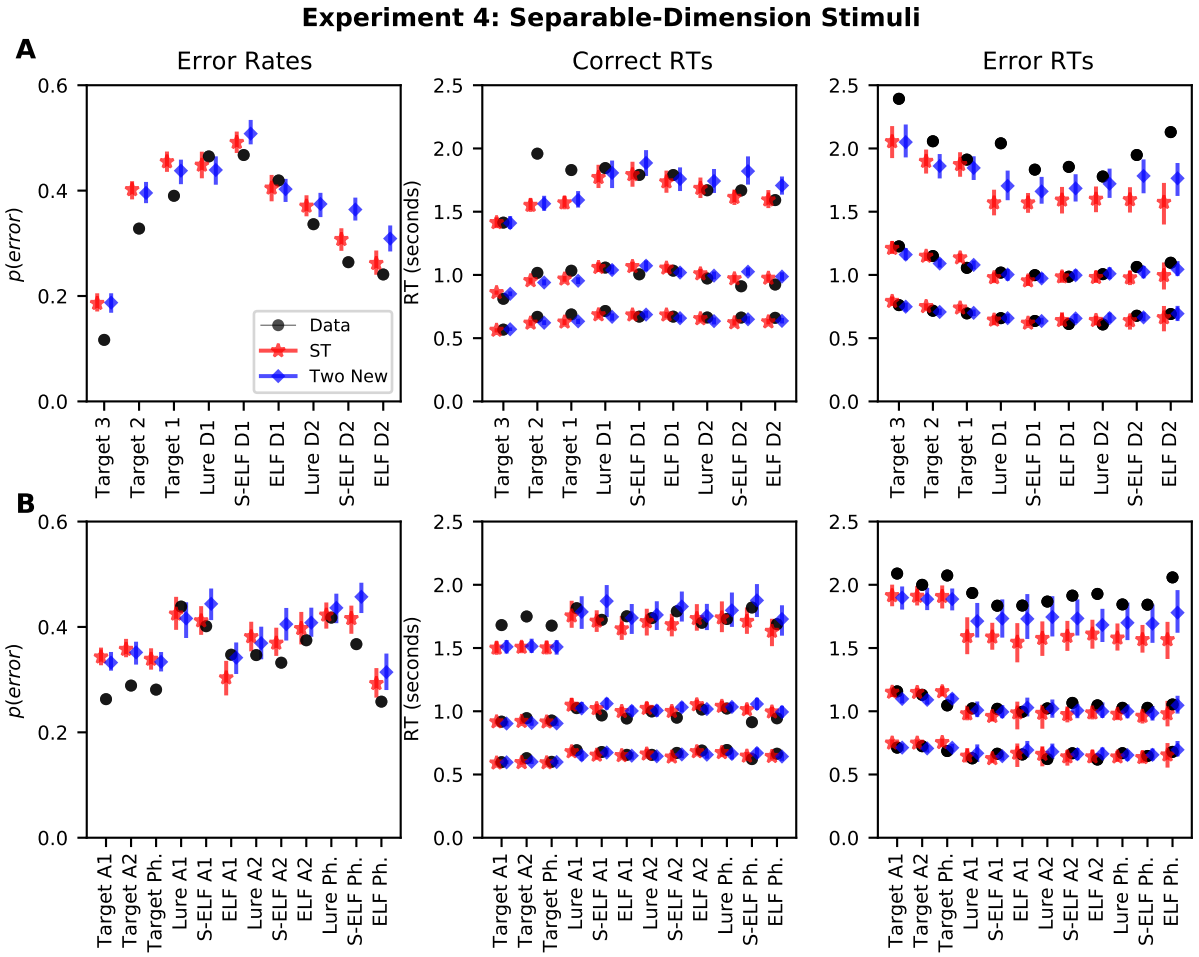

Figure 4. Group-averaged error rates (left panel) and correct and error RT distributions (middle and right panels, respectively) for the data from Experiment 4 with separable-dimension stimuli. RT distributions are summarized using the .1, .5, and .9 quantiles. The top row (A) shows targets separated by serial position and lure results separated by distance. The bottom row (B) shows the results separately for each fixed dimension in the probe stimulus. Model predictions are group-averaged posterior predictives from the parallel EB-LBA models with different decision rules, including the model with a self-terminating rule for "new" responses (red) and the model where two "new" accumulators are required to hit the threshold (blue). Error bars depict the 95% highest density interval (HDI). Note: D1 = distance 1, D2 = distance 2, H = height, Sat. = saturation, BP = bar position.

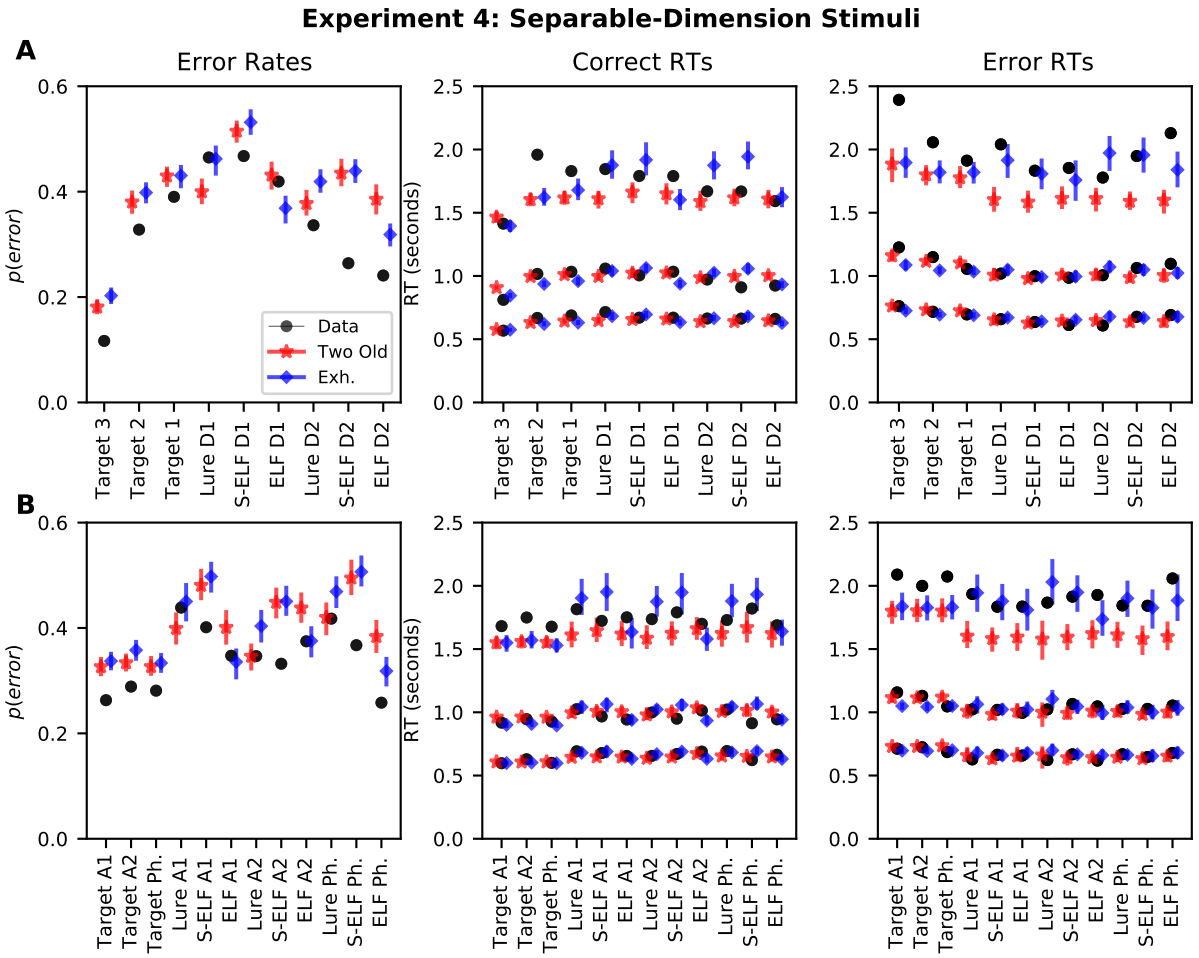

Figure 5. Group-averaged error rates (left panel) and correct and error RT distributions (middle and right panels, respectively) for the data from Experiment 4 with separable-dimension stimuli. RT distributions are summarized using the .1, .5, and .9 quantiles. The top row (A) shows targets separated by serial position and lure results separated by distance. The bottom row (B) shows the results separately for each fixed dimension in the probe stimulus. Model predictions are group-averaged posterior predictives from the parallel EB-LBA models with different decision rules, including the model where two "old" accumulators are required to hit threshold (red) and the exhaustive model (blue). Error bars depict the 95% highest density interval (HDI). Note: D1 = distance 1, D2 =f distance 2, H = height, Sat. = saturation, BP = bar position.

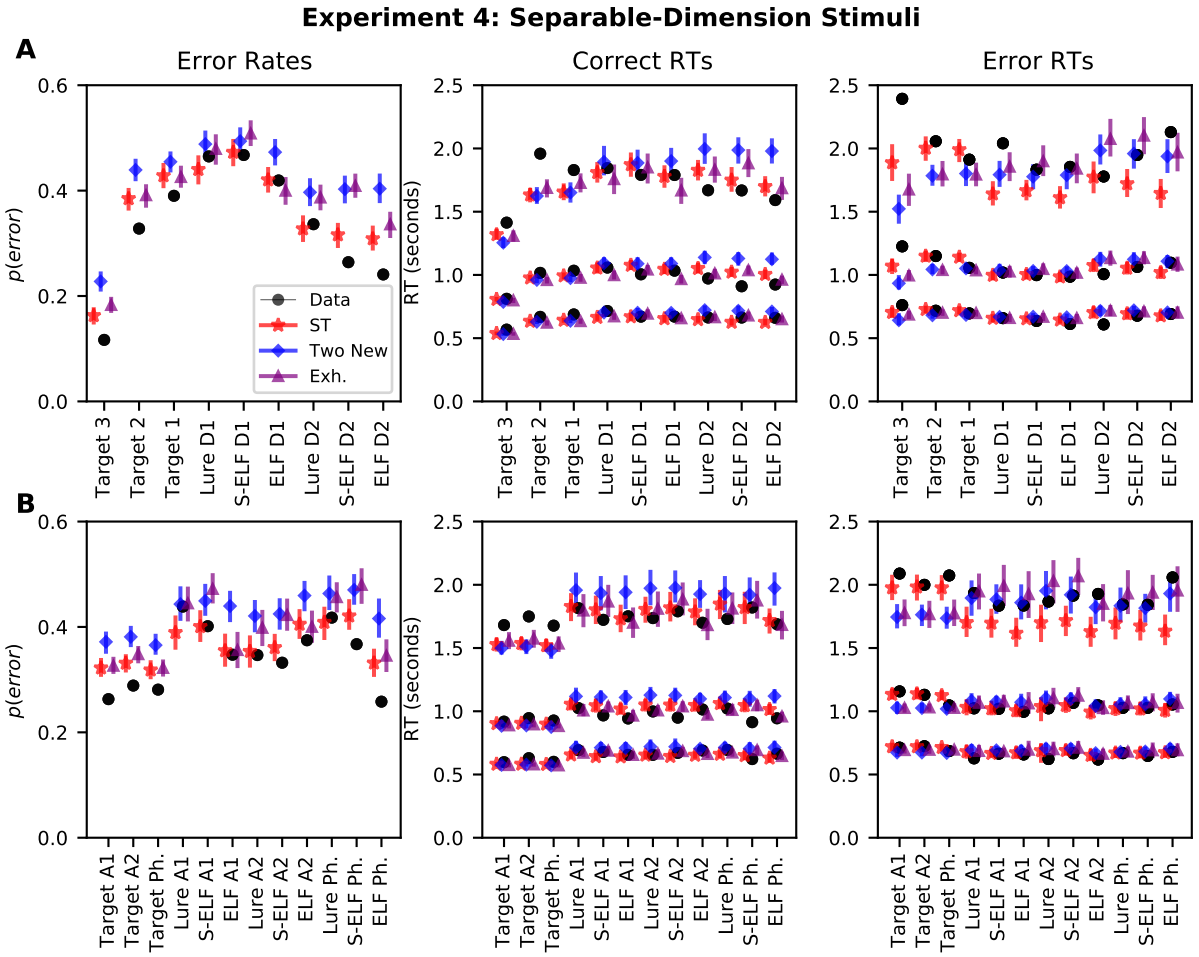

Figure 6. Group-averaged error rates (left panel) and correct and error RT distributions (middle and right panels, respectively) for the data from Experiment 4 with separable-dimension stimuli. RT distributions are summarized using the .1, .5, and .9 quantiles. The top row (A) shows targets separated by serial position and lure results separated by distance. The bottom row (B) shows the results separately for each fixed dimension in the probe stimulus. Model predictions are group-averaged posterior predictives from the hybrid coactive-parallel EB-LBA models with different decision rules for "new" responses, including the self-terminating model (red), the model where two "new" accumulators are required to hit the threshold (blue), and the exhaustive model (purple). Error bars depict the 95% highest density interval (HDI). Note: D1 = distance 1, D2 = distance 2, H = height, Sat. = saturation, BP = bar position.

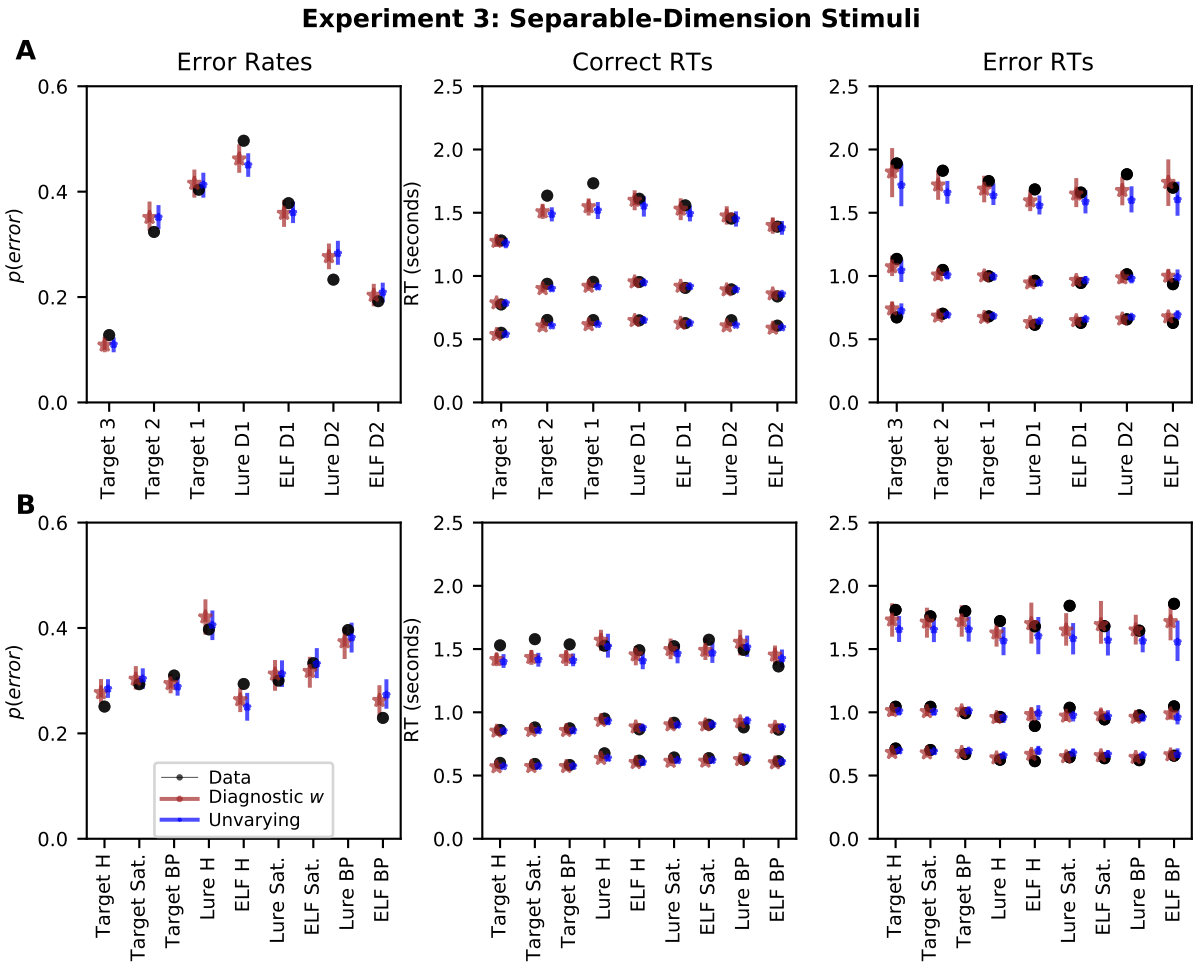

Figure 7. Group-averaged error rates (left panel) and correct and error RT distributions (middle and right panels, respectively) for the data from Experiment 3 with separable-dimension stimuli. RT distributions are summarized using the .1, .5, and .9 quantiles. The top row (A) shows targets separated by serial position and lure results separated by distance. The bottom row (B) shows the results separately for each fixed dimension in the probe stimulus. Model predictions are group-averaged posterior predictives from the diagnostic attention model (brown, referred to as "diagnostic w" in the legend) and the attention to unvarying dimensions model (referred to as "unvarying"). Error bars depict the 95% highest density interval (HDI). Note: D1 = distance 1, D2 = distance 2, H = height, Sat. = saturation, BP = bar position.

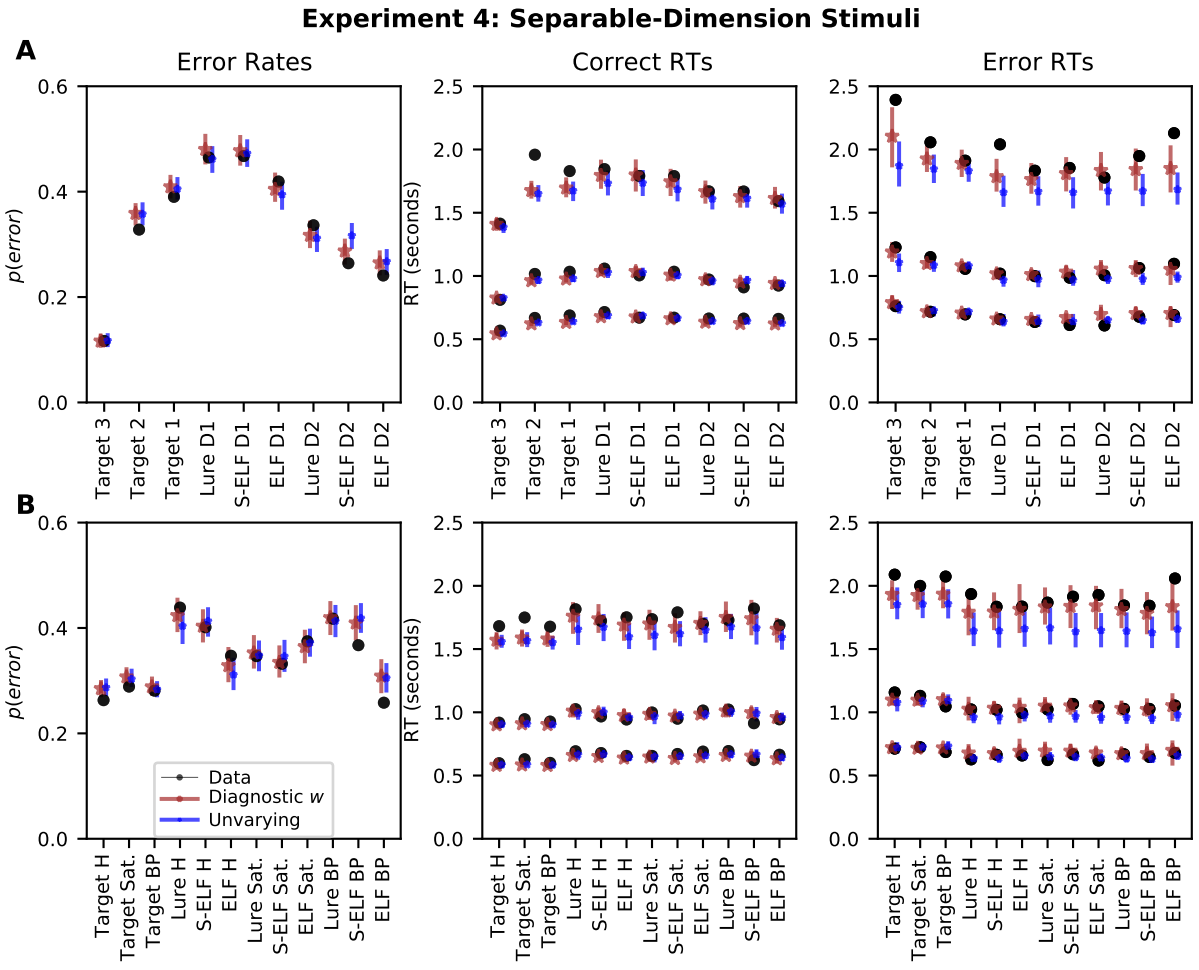

Figure 8. Group-averaged error rates (left panel) and correct and error RT distributions (middle and right panels, respectively) for data from Experiment 4 with separable-dimension stimuli. RT distributions are summarized using the .1, .5, and .9 quantiles. The top row (A) shows targets separated by serial position and lure results separated by distance. The bottom row (B) shows results separately for each fixed dimension in the probe stimulus. Model predictions are group-averaged posterior predictives from the diagnostic attention model (brown, referred to as "diagnostic w" in the legend) and the attention to unvarying dimensions model (referred to as "unvarying"). Error bars depict the 95% highest density interval (HDI). Note: D1 = distance 1, D2 = distance 2, H = height, Sat. = saturation, BP = bar position.

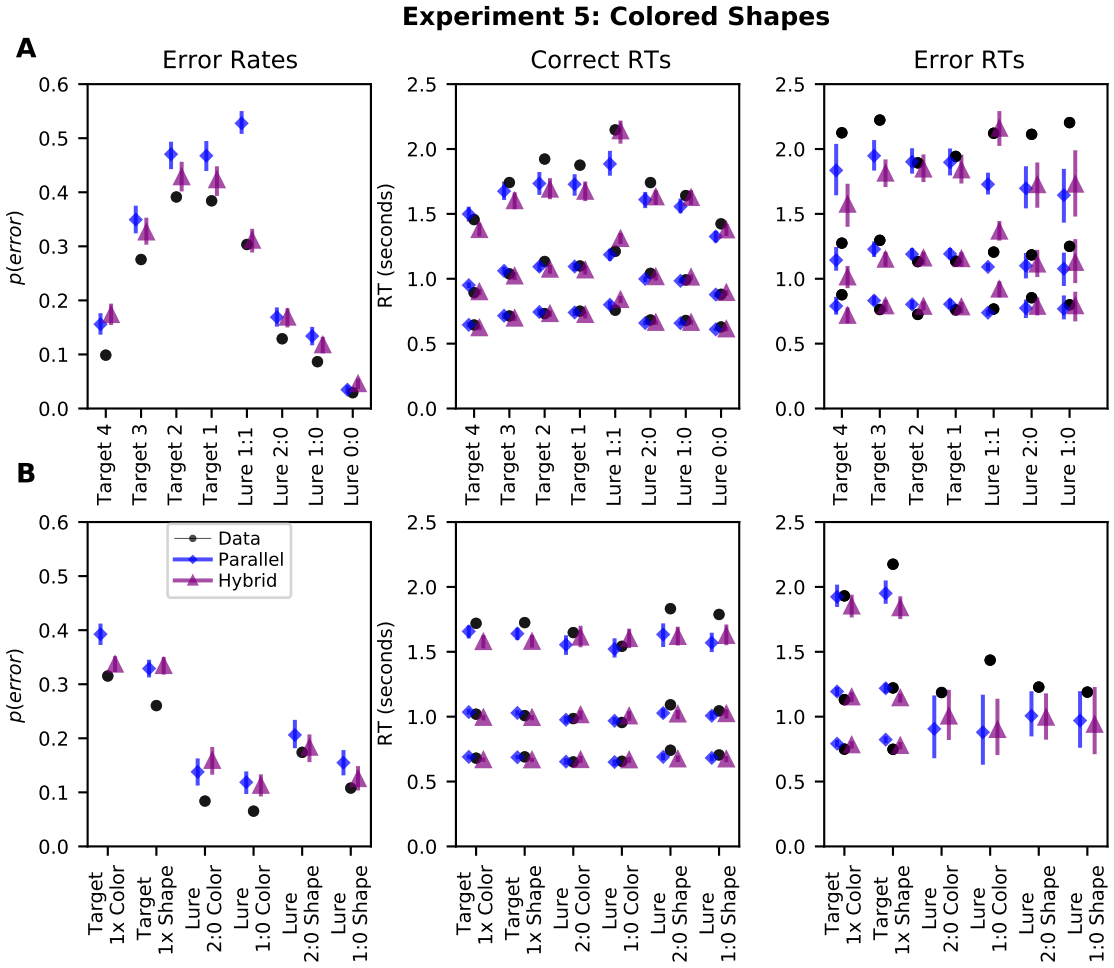

Figure 9. Group-averaged choice probabilities (left panel) and correct and error RT distributions (middle and right panels, respectively) for the data from Experiment 5 with colored shapes as stimuli along with posterior predictives from the parallel model and the hybrid coactive-parallel EB-LBA model. RT distributions are summarized using the .1, .5, and .9 quantiles. The top row (A) shows results for target (broken down by serial position) and lure (1:1, 2:0, 1:0, and 0:0) trials. The bottom row (B) breaks down results for targets depending on whether the shape or color is the once-presented feature along with 1:0 and 2:0 probes when color or shape is the extralist feature. Error bars depict the 95% highest density interval (HDI).