A global matching model of choice and response times in the Deese-Roediger-McDermott semantic and structural false recognition paradigms: Supplementary Materials

Adam F. Osth, Lyulei Zhang, & Samuel Williams

The University of Melbourne

Address correspondence to:

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

Author Note

.

A global matching model of choice and response times in the Deese-Roediger-McDermott semantic and structural false recognition paradigms: Supplementary Materials

A: Simulations of Cross-Task Correlations in Experiment 3 with Variability in w

In the main text, we described a result from Ballou and Sommers (2008) that evaluated relationships between the semantic and structural DRM tasks at the level of individual participants. Specfically, they calculated the correlations for true and false recognition across the semantic and structural DRM tasks at the level of individual participants. They found a significant and positive correlation for true recognition (e.g., hit rates between the two tasks) but a small and non-significant correlation for false recognition (e.g., false alarm rates to critical lures from studied categories across the two tasks). Based on this result, they concluded the two tasks were independent of each other.

How can a single process model like ours produce such different correlations for true and false recognition? As mentioned in the main text, the w parameter provides a partial explanation for this. As mentioned in the main text, as w approaches 1, the similarity between a probe and a trace become dominated by semantic similarity, and thus we should see strong semantic false recognition but no structural false recognition. However, as w approaches 0, similarity is dominated by structural similarity and we should see strong structural false recognition. If variability in w is high, false recognition will become de-correlated across tasks because different participants will show very different false alarm rates in each task.

Critically, the variability in w will have little effect on the predicted hit rates because w does not affect the self-similarity (the similarity between a probe and its own representation). It will still have some effect because hit rate predictions are also affected by the similarity between the probe and other items in the memory set, but these are all relatively small contributions compared to the self similarity.

In order to demonstrate the effect of variability of w on the cross-task correlations, we simulated the paradigms of Experiment 3A and 3B patterns using 50 randomly-sampled

sets of participant parameters from Experiments 1 and 2. We used these older parameter sets to indicate that the behavior of the model is not dependent on its contact with data from Experiment 3. However, we substituted different values of w for each participant and manipulated this variability across four levels – for each level, we simulated the hit and false alarm rates for Experiment 3. Specifically, we generated values of w from a truncated normal distribution bounded between 0 and 1 with a mean of .50 and four different standard deviations: .01, .15, .4, and .75.

The main text details the procedures for Experiment 3. An important difference between the two is that in Experiment 3A, participants study 9 exemplars from each category and are tested on 3 targets and 1 critical lure from each category. The fact that three times as many targets are tested as critical lures is consistent with the Ballou and Sommers paradigm. In Experiment 3B, participants study 11 exemplars from each category and are tested on 1 target and 1 critical lure from each category. In each experiment, participants are tested on a total of 60 critical lures per task per participant.

The simulation results can be seen in Figure 1 for both target trials (true recognition) and critical lure trials (false recognition). The x-axis contains the predicted endorsement rates from the structural DRM task while the y-axis contains the predicted endorsement probabilities for the semantic DRM task. When w^{σ} is low (.01) the correlation between the two tasks is high for both true and false recognition. However, as w^{σ} is increased (from left to right), the correlation between false recognition of the two tasks decreases much more radically than for true recognition.

As we mention in the main text, variability in w is only one of many factors that can affect the cross-task correlations. The size of the dataset is another – our datasets test five times as many critical lures as those from Ballou and Sommers, and we would expect much smaller correlations under those conditions. In addition, other parameters such as differences in encoding strength across the two tasks can have a further influence.

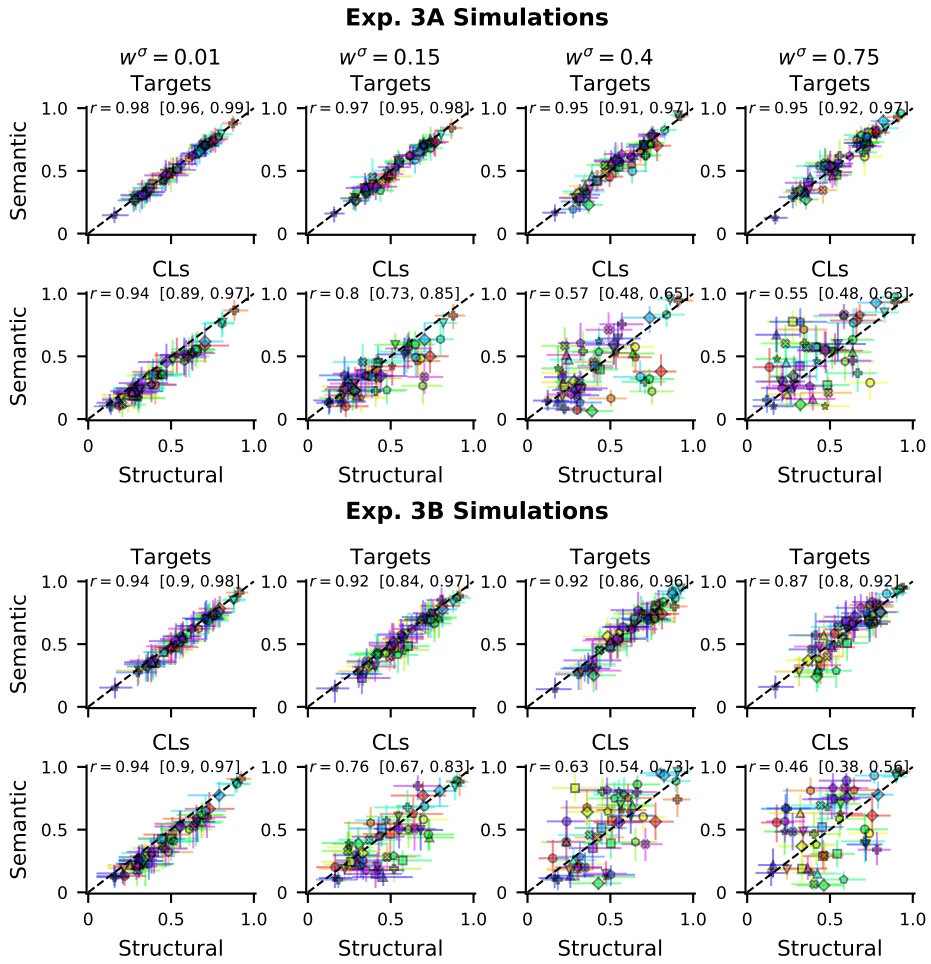

Figure 1. Simulations of the relationship between semantic and structural false recognition in Experiments 3A and 3B with four different standard deviations of w. See the text for more details. Note that CLs = critical lures.

B: Experiment 3 Fits to Individual Participants

To preserve space in the main text, we refrained from showing the fits to individual participants and opted to show them here instead. Figure 2 shows the fits to individual participant hit and false alarm rates from both the semantic (left) and structural (right) DRM tasks. One can see that the model continues to provide an excellent account of individual participants. The poorest is the coverage of the hit rates in the structural DRM task in Experiment 3B ($r = .75$). This may be an anomaly because Experiment 3A – which is very similar – shows a strong account of individual participant hit rates in the structural

DRM task (r = .92).

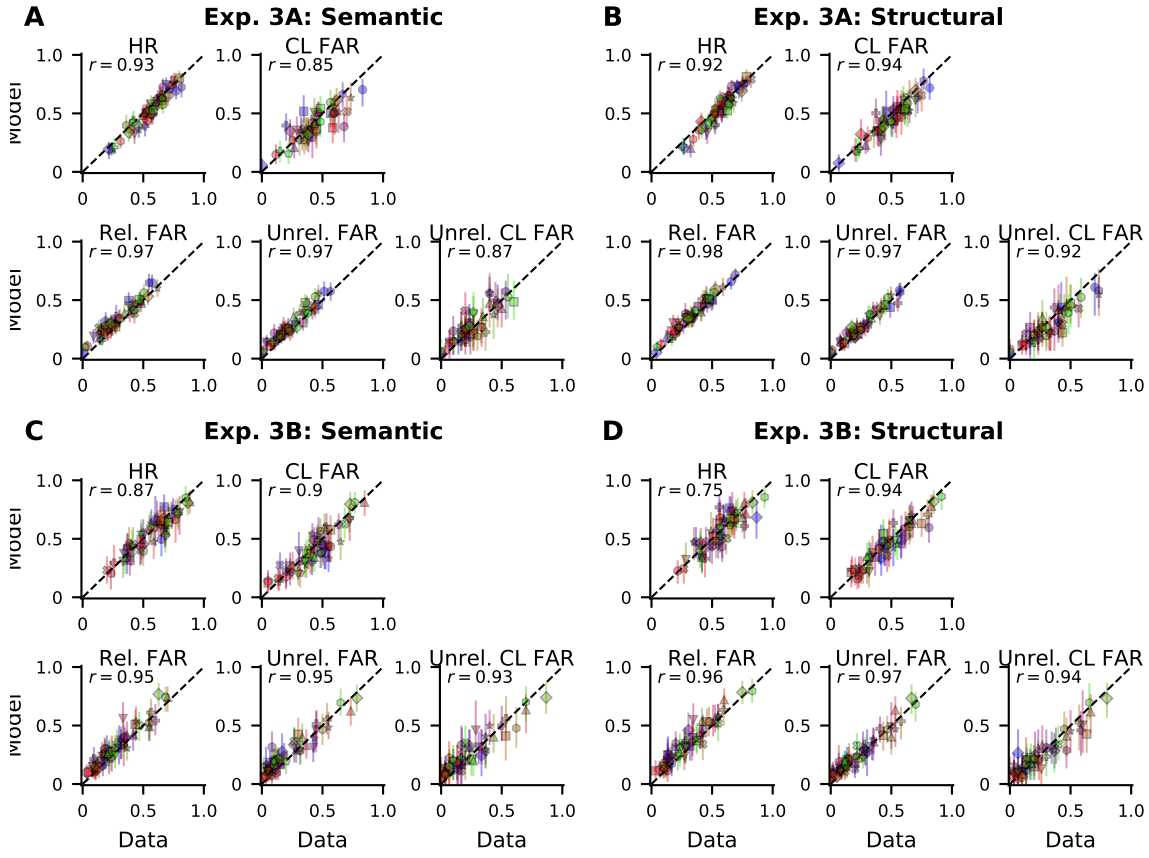

Figure 2. Individual participant hit and false alarm rates for each trial type from both the data and the model from Experiment 3. Note that HR = hit rate, FAR = false alarm rate, CL = critical lure, rel. = related lure, unrel. = unrelated lure, and unrel. CL = unrelated critical lure. Error bars represent 95% highest density intervals (HDIs).

Individual participant RT distributions from the data and the model can be seen in Figure 3 for Experiment 3A and Figure 4 for Experiment 3B. As mentioned in the main text, RT distributions are summarized using the .1, .5, and .9 quantiles. These figures reveal that the model continues to provide an excellent account of cross-participant variation in RTs. The weakest correspondence between the model and the data is in the places that are expected, namely the .9 quantile, which is the most variable region of the RT distribution.

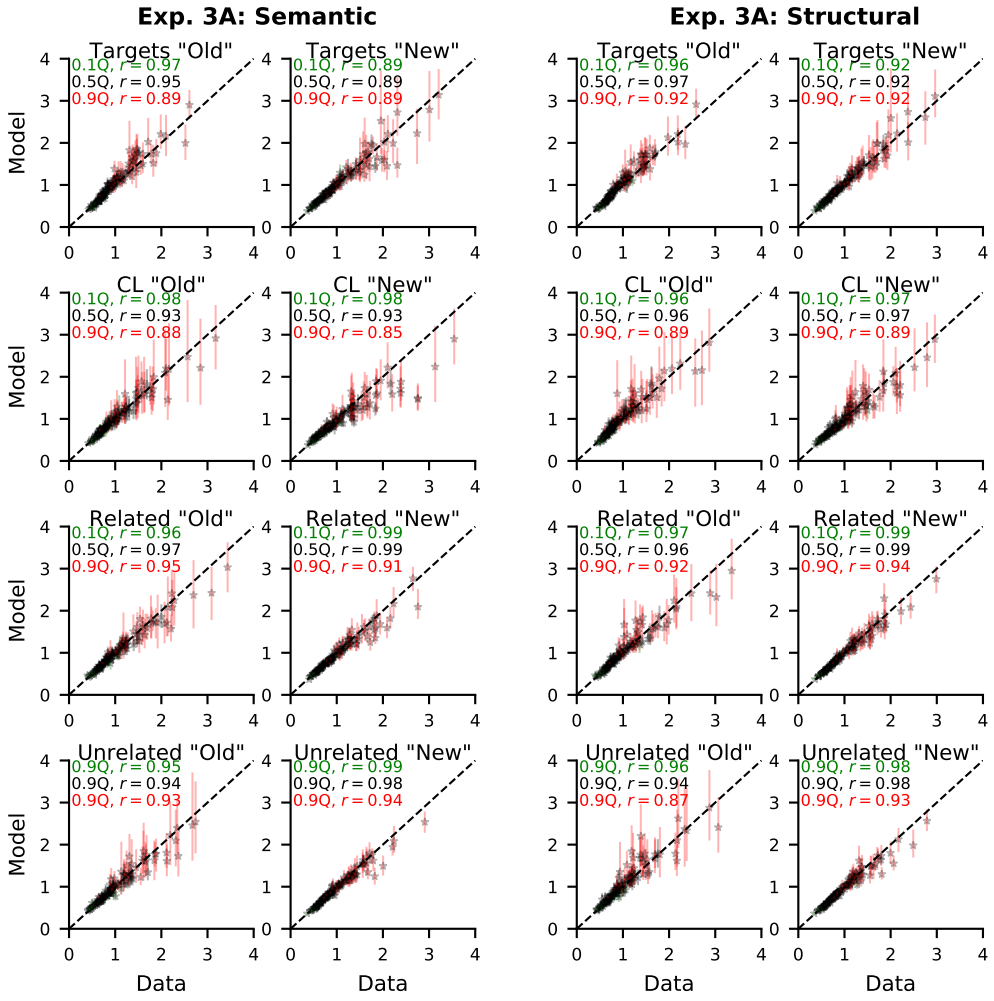

Figure 3. Individual participant RT distributions (summarized using the .1, .5, and .9 quantile) for each trial type from both the data and the model from Experiment 3A. Note that CL = critical lure, rel. = related lure, unrel. = unrelated lure, and unrel. CL = unrelated critical lure. Error bars represent 95% highest density intervals (HDIs).

Semantic-Structural Correlations for Other Lure Types

The main text focuses on the correlations between the semantic and structural DRM paradigms for targets and critical lures. Ballou and Sommers (2008) additionally reported the correlations for related lures and found that they strongly resembled those of target items. For this reason, we additionally analyzed the correlations for all of our other lure types, namely related lures, unrelated lures, and unrelated critical lures.

Scatterplots for both the data and the model can be seen in Figure 5, which

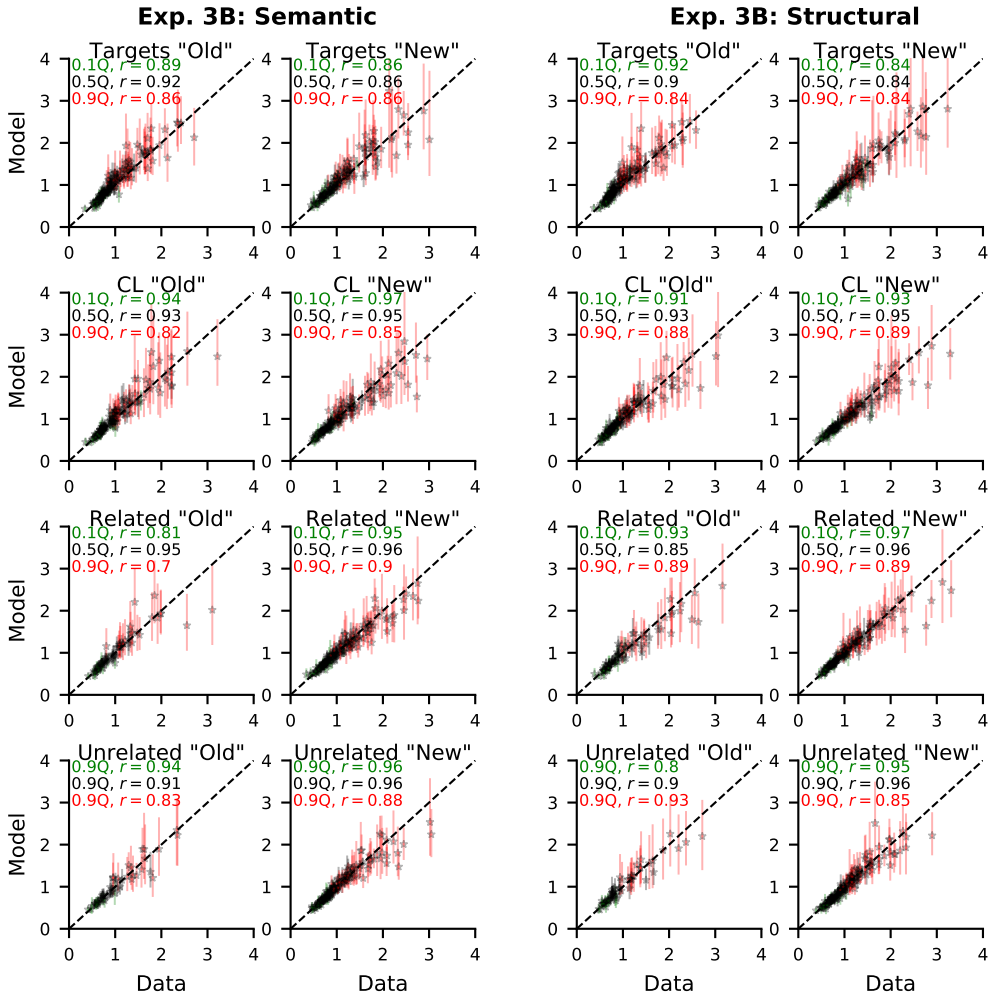

Figure 4. Individual participant RT distributions (summarized using the .1, .5, and .9 quantile) for each trial type from both the data and the model from Experiment 3B. Note that CL = critical lure, rel. = related lure, unrel. = unrelated lure, and unrel. CL = unrelated critical lure. Error bars represent 95% highest density intervals (HDIs).

separately reports the results from Experiment 3A (column A, left) and 3B (column B, right). Our results strongly resemble those of Ballou and Sommers - the correlations for each of the other lure types are high and strongly resemble those of target items (e.g., r˜.9). The one exception is that unrelated critical lures have a low correlation in Experiment 3A (r = .73), but was "in the ballpark" of the other correlation estimates in Experiment 3B (r = .89). Critically, the model reproduces each of these correlations quite well.

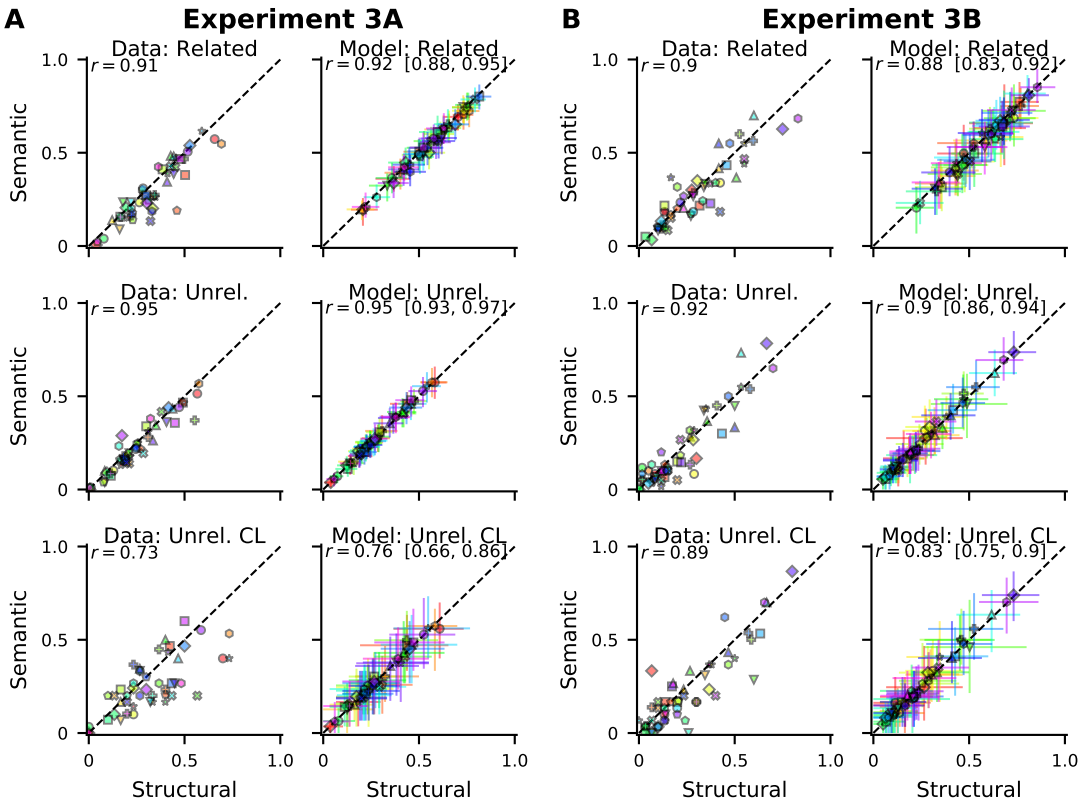

Figure 5. Scatterplots showing the relationship between the semantic (x-axis) and structural (y-axis) DRM task in Experiment 3A (column A, left) and 3B (column B, right). This figure shows the results for related lures (top row), unrelated lure (middle row) and unrelated critical lures (bottom row). For the model, we reported both the mean and 95% highest density interval (HDI) of the correlation coefficient. Error bars show the 95% HDIs on the model's hit and false alarm rate predictions.

C: Varied Encoding Strength Across Semantic and Structural Lists in Experiment 3B

In the main text, we found that our model over-predicted the cross-task correlations for both true and false recognition in Experiment 3B. The model we employed assumes that encoding strength for both semantic and structural lists is the same, which may be an overly restrictive assumption. Here, we explored whether relaxing such assumptions gave a stronger account of the cross-task correlations.

We allowed w, m, and τ to vary across the semantic and structural DRM tasks. This model contained 11 parameters per participant, while the core model only contained 10

parameters per participant. This model resulted in slightly worse memory strength in the structural DRM task (mean $m_{\text{structural}^{\mu}} = .90$), more semantic encoding in the semantic DRM task (mean $w^{\mu}_{\text{semantic}} = .62$) than in the structural DRM task ($w^{\mu}_{\text{structural}} = .48$), as well as higher precision in the semantic DRM task (mean $\tau^{\mu}_{\text{semantic}} = 2.06$) than in the structural DRM task (mean $\tau^{\mu}_{\text{structural}} = 1.86$). The scatterplots showing the predicted hit and false alarm rates for each task can be seen in Figure 6. These reveal that with some defficient encoding in the structural DRM task, the model's correlations are within the range of those seen in the data.

We would like to emphasize that we are *not* making strong claims about encoding strengths varying across the two tasks. Our materials are different in each task and it may well be the case that the structural task contains more difficult words, and that this difficulty is not well captured by our structural or semantic representations.

Nonetheless, an encoding explanation is not entirely implausible. In Experiment 3B, participants are presented with 11 exemplars per category. The structural categories are different from the semantic ones because the perceived homogeneity is very high – the words share all but one or two letters. The exemplars in the semantic categories all bear strong associative relationships with the critical lure but not necessarily with each other, so the relationships may not be as obvious to participants. When perceived homogeneity is high, it may be the case that encoding gets worse because either a.) participants are less interested in viewing a high degree of obviously similar words or b.) similar content may produce worse encoding overall, as predicted by similarity-sensitive encoding models (Farrell & Lewandowsky, 2002; Elhalal, Davelaar, & Usher, 2014).

D: Experiment 4 Fits to Individual Participants

To preserve space in the main text, we show fits to individual participants from Experiment 4 here. Experiment 4 utilized a 2x2 factorial design which manipulated depth-of-processing (deep vs. shallow) and DRM task (semantic vs. structural). Fits to the

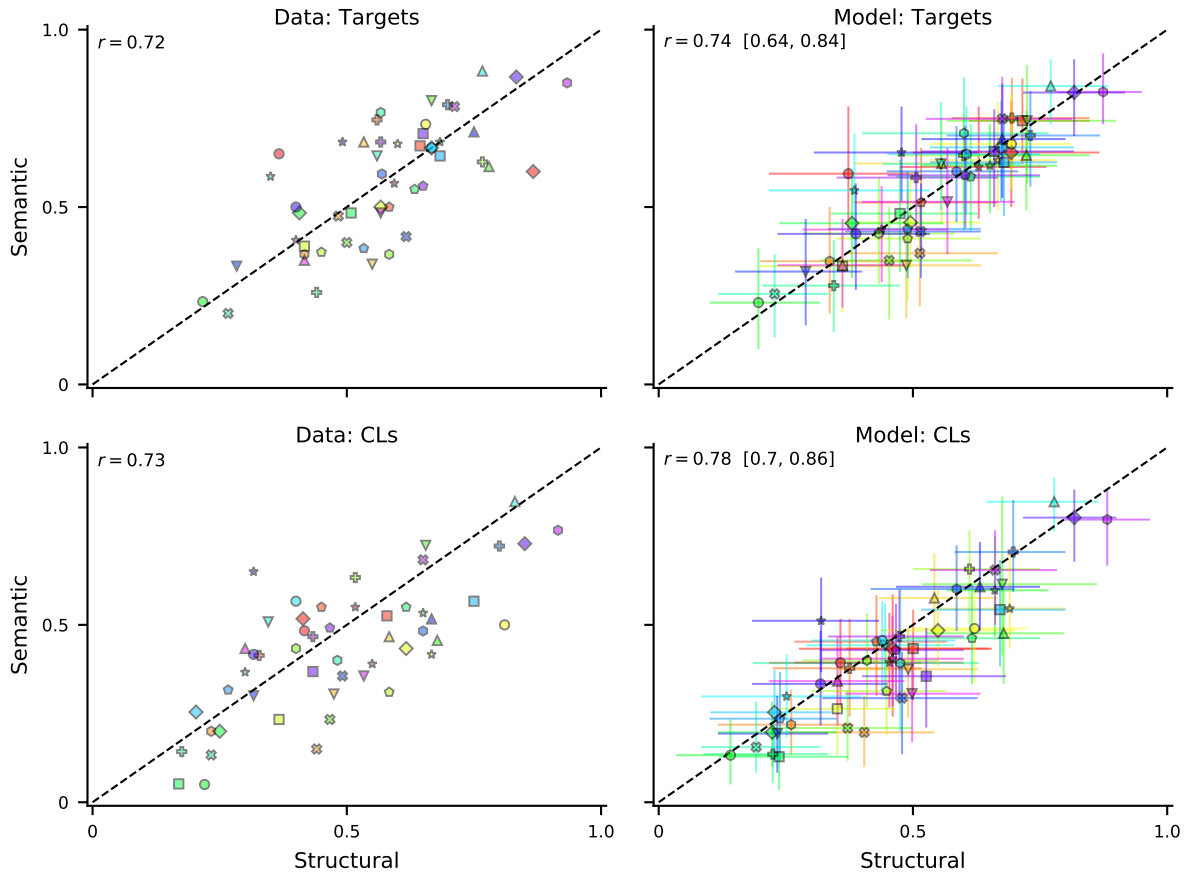

Figure 6. Scatterplots showing the relationship between the semantic (x-axis) and structural (y-axis) DRM task in Experiment 3B. For each task, we compared the hit rate to target items (top) and the false alarm rate to critical lures from studied categories (bottom). For the model, we reported both the mean and 95% highest density interval (HDI) of the correlation coefficient. Error bars show the 95% HDIs on the model's hit and false alarm rate predictions.

hit and false alarm rates from each trial type can be seen in Figure 7. These reveal that the model provides an excellent account of these individual differences. The fact that the model does so well with capturing individual variation across individual participants in the false alarm rates to critical lures is surprising because there are only 30 critical lures in each cell of the design.

Figure 8 shows the fits to individual participant RT distributions for "old" and "new" responses. Because of the sparsity of the data in the 2x2 design, we collapsed across both

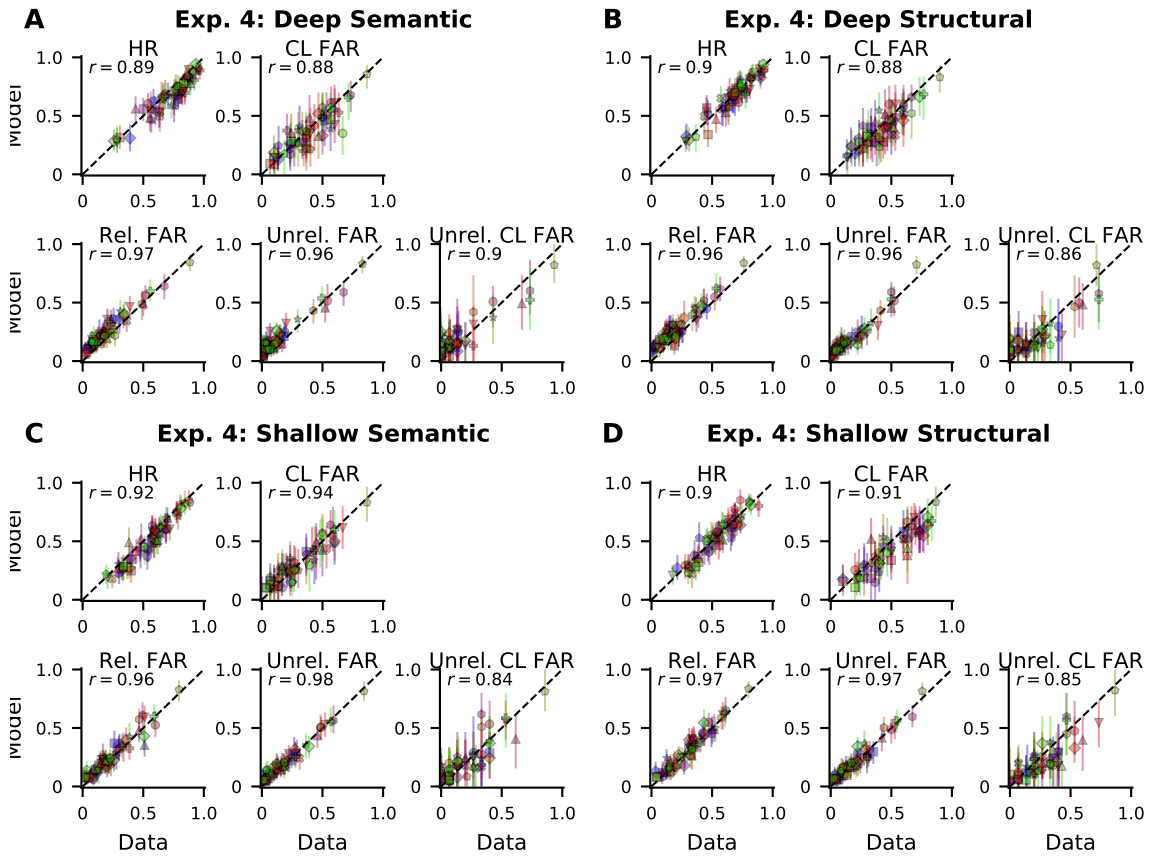

Figure 7. Individual participant hit and false alarm rates for each trial type from both the data and the model from Experiment 4. Note that HR = hit rate, FAR = false alarm rate, CL = critical lure, rel. = related lure, unrel. = unrelated lure, and unrel. CL = unrelated critical lure. Error bars represent 95% highest density intervals (HDIs).

the semantic and structural DRM tasks here to focus on the difference between deep vs. shallow processing. Once again, the model continues to provide a very strong account of individual differences in RTs.

Cross-Task Correlations

Figure 9 shows the cross-task correlations for true and false recognition in Experiment 4, which were conducted separately for both the deep and shallow processing conditions. Interestingly, false recognition under deep processing showed a smaller cross-task correlation ($r = .56$) than shallow processing ($r = .69$). The model largely

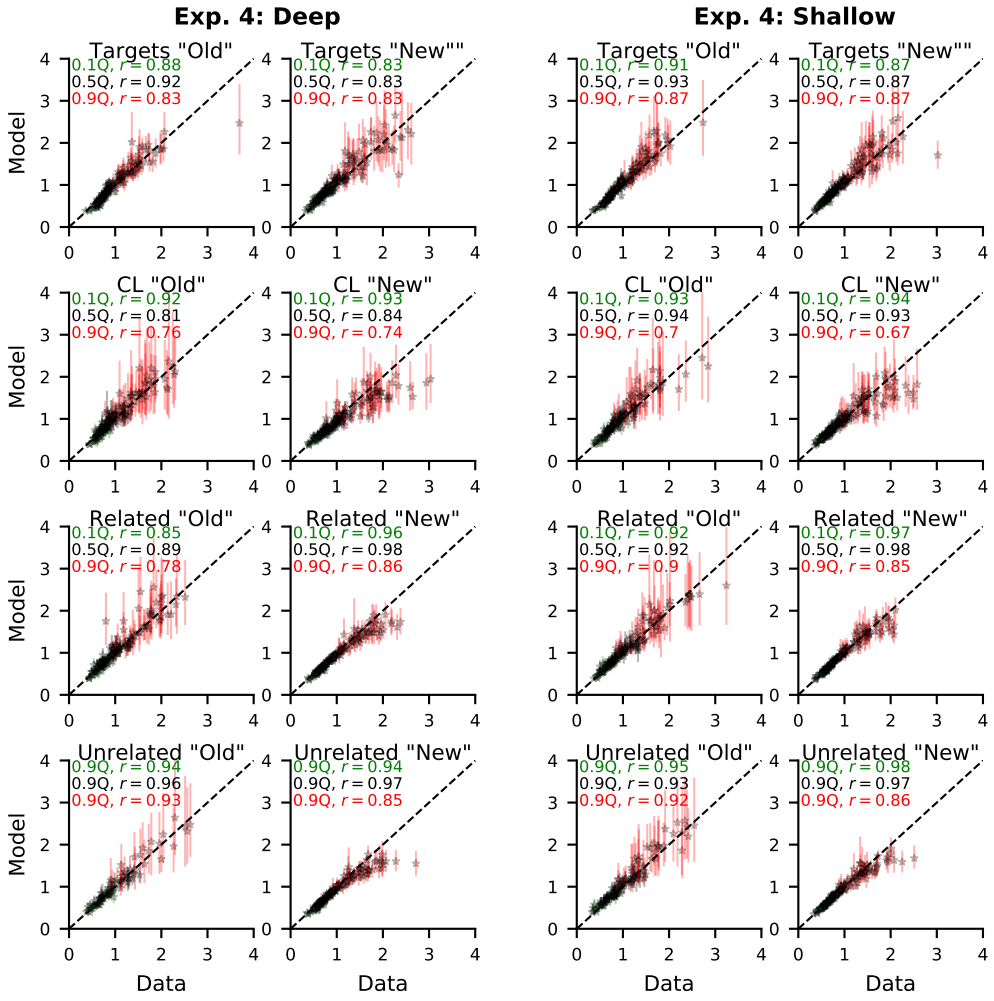

Figure 8. Individual participant RT distributions (summarized using the .1, .5, and .9 quantile) for each trial type from both the data and the model from Experiment 4. These results were collapsed across both the semantic and structural DRM tasks. Note that CL = critical lure, rel. = related lure, unrel. = unrelated lure, and unrel. CL = unrelated critical lure. Error bars represent 95% highest density intervals (HDIs).

reproduced these correlations. The cross-task correlations for false recognition were all within the 95% HDI of the model. The model predicted slightly higher correlations for true recognition than seen in the data but the differences were numerically quite small.

E: Experiment 5 Fits to Individual Participants

To preserve space in the main text, we show fits to individual participants from Experiment 5 here. Experiment 5 utilized a 2x2 factorial design which manipulated study

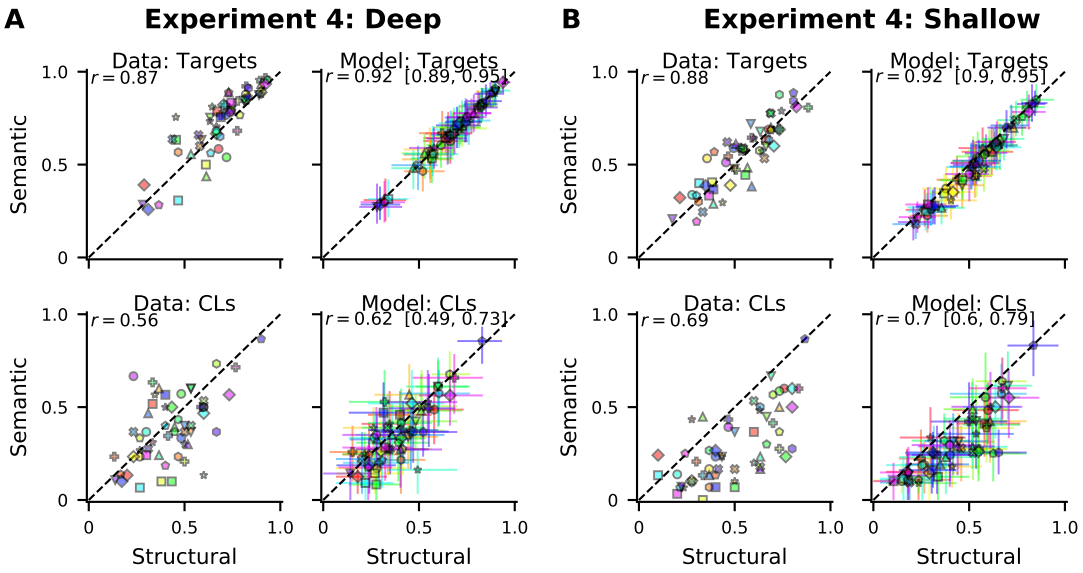

Figure 9. Scatterplots showing the relationship between the semantic (x-axis) and structural (y-axis) DRM tasks for the deep and shallow processing conditions in Experiment 4. For each task, we compared the hit rate to target items (top) and the false alarm rate to critical lures from studied categories (bottom). For the model, we reported both the mean and 95% highest density interval (HDI) of the correlation coefficient. Error bars show the 95% HDIs on the model's hit and false alarm rate predictions.

time (long vs. short) and DRM task (semantic vs. structural). Fits to the hit and false alarm rates from each trial type can be seen in Figure 12. These reveal that the model provides an excellent account of these individual differences.

Figure 13 shows the fits to individual participant RT distributions for "old" and "new" responses. Because of the sparsity of the data in the 2x2 design, we collapsed across both the semantic and structural DRM tasks here to focus on the difference between long vs. short study time conditions. Once again, the model continues to provide a very strong account of individual differences in RTs.

Cross-Task Correlations

Figure 11 shows the cross-task correlations for true and false recognition in Experiment 5, which were conducted separately for both the long and shor study tiem conditions. Interestingly, similar to Experiment 4, false recognition in the condition that

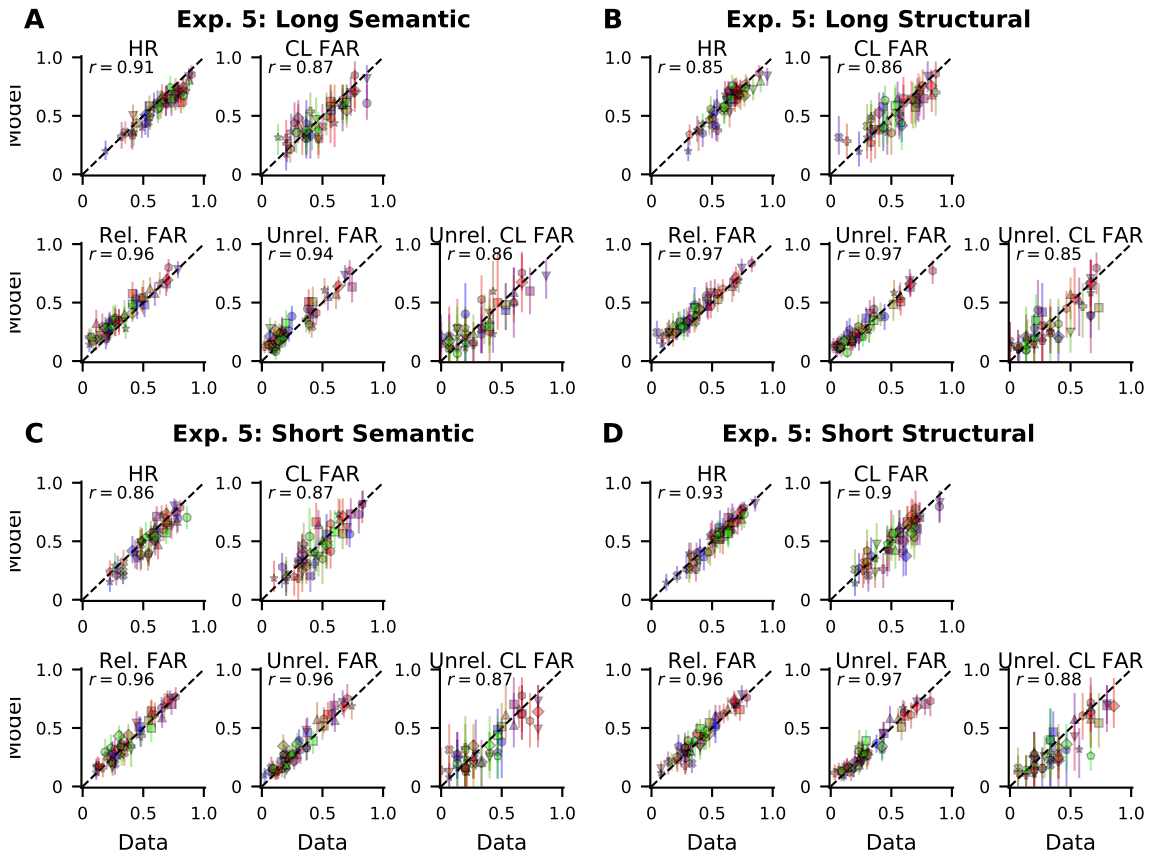

Figure 10. Individual participant hit and false alarm rates for each trial type from both the data and the model from Experiment 5. Note that HR = hit rate, FAR = false alarm rate, CL = critical lure, rel. = related lure, unrel. = unrelated lure, and unrel. CL = unrelated critical lure. Error bars represent 95% highest density intervals (HDIs).

showed better performance (the long study time condition) showed a smaller cross-task correlation ($r = .61$) than the condition with worse performance (the shor study time condition: $r = .78$). The model largely reproduced these correlations. The cross-task correlations for false recognition were all within the 95% HDI of the model, and overall the model reproduces the correlations quite well.

F: Experiment 6 Fits to Individual Participants

To preserve space in the main text, we show fits to individual participants from Experiment 6 here. Experiment 6 utilized a 2x2 factorial design which manipulated time

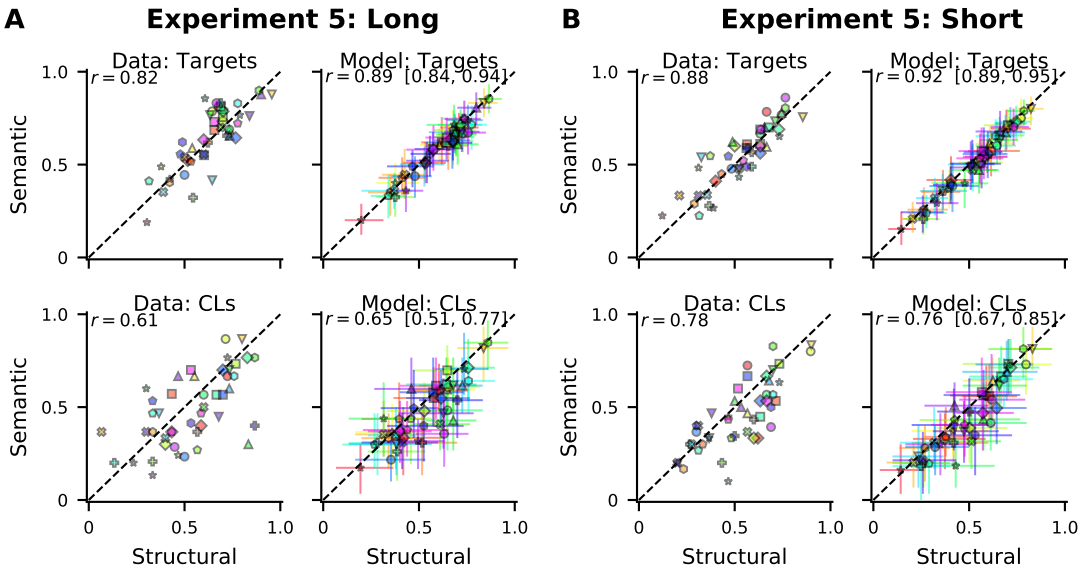

Figure 11. Scatterplots showing the relationship between the semantic (x-axis) and structural (y-axis) DRM tasks for the long and short study time conditions in Experiment 5. For each task, we compared the hit rate to target items (top) and the false alarm rate to critical lures from studied categories (bottom). For the model, we reported both the mean and 95% highest density interval (HDI) of the correlation coefficient. Error bars show the 95% HDIs on the model's hit and false alarm rate predictions.

pressure (speed- vs. accuracy-emphasis) and DRM task (semantic vs. structural). We focused on the augmented model here, which was the winner in model selection. The augmented model deviated from the core model because both the w (relative weight of semantic and structural representations) and τ (memory precision) parameters were allowed to vary across the time pressure conditions. This model contained 14 parameters per participant.

Fits to the hit and false alarm rates from each trial type can be seen in Figure **??**. These reveal that the model provides a reasonable account of the individual differences in this experiment, although the model's ability to capture these differences is somewhat poorer in this experiment than the others. Notably, for the structural DRM task in the accuracy-emphasis condition, the correspondence between the model's critical lure FAR and those in the data was good, but not nearly as strong as in the previous experiments ($r = .64$). However, the model provides a very reasonable account of the same task in the

speed-emphasis condition (r = .87).

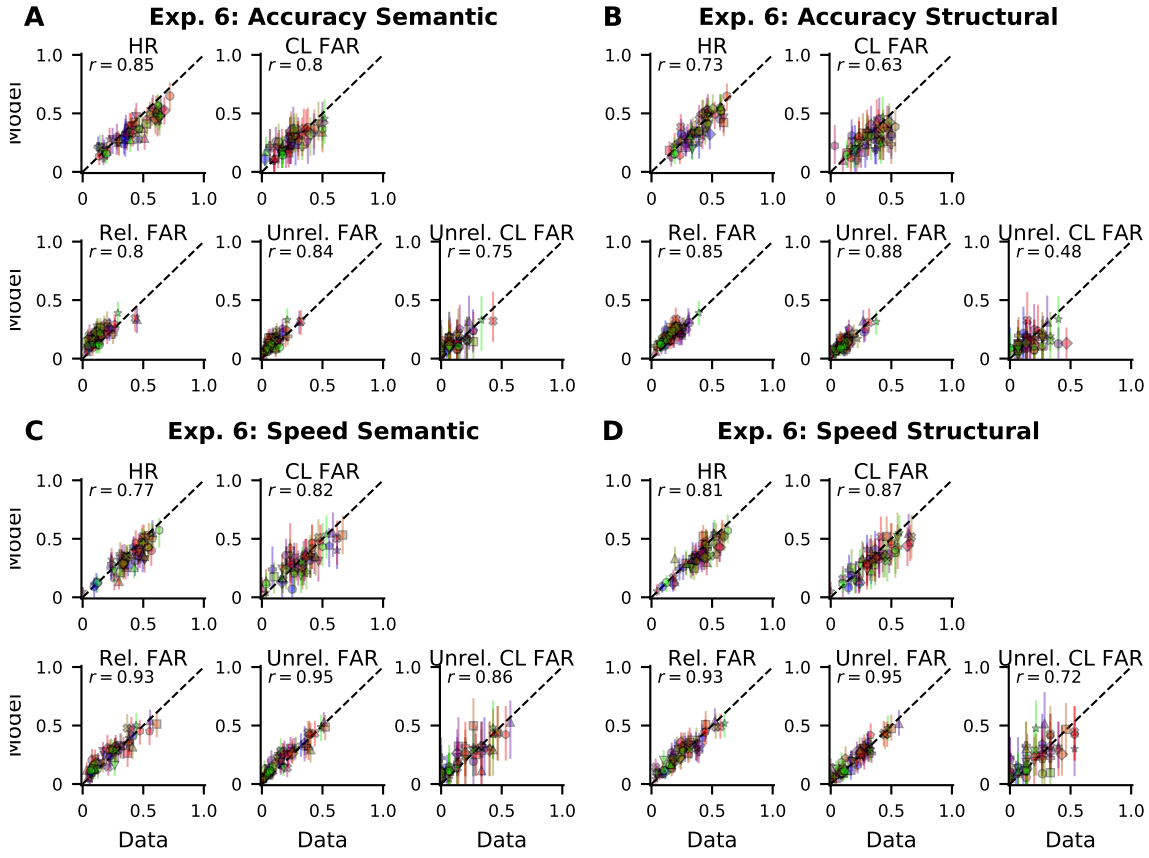

Figure 12. Individual participant hit and false alarm rates for each trial type from both the data and the model from Experiment 6. Note that HR = hit rate, FAR = false alarm rate, CL = critical lure, rel. = related lure, unrel. = unrelated lure, and unrel. CL = unrelated critical lure. Error bars represent 95% highest density intervals (HDIs).

The model's fits to the individual participant RT data can be seen in Figure **??**, which collapses across the semantic and structural DRM tasks due to the low amount of data. Again, the model provides a very good account of the individual participant data, but this account is not quite as strong as in previous experiments. Nonetheless, there were very few systematic misses in this experiment. While there are participants who exhibit .9 quantiles that are slower in the data than the model predicts, this appears in only a small number of cases.

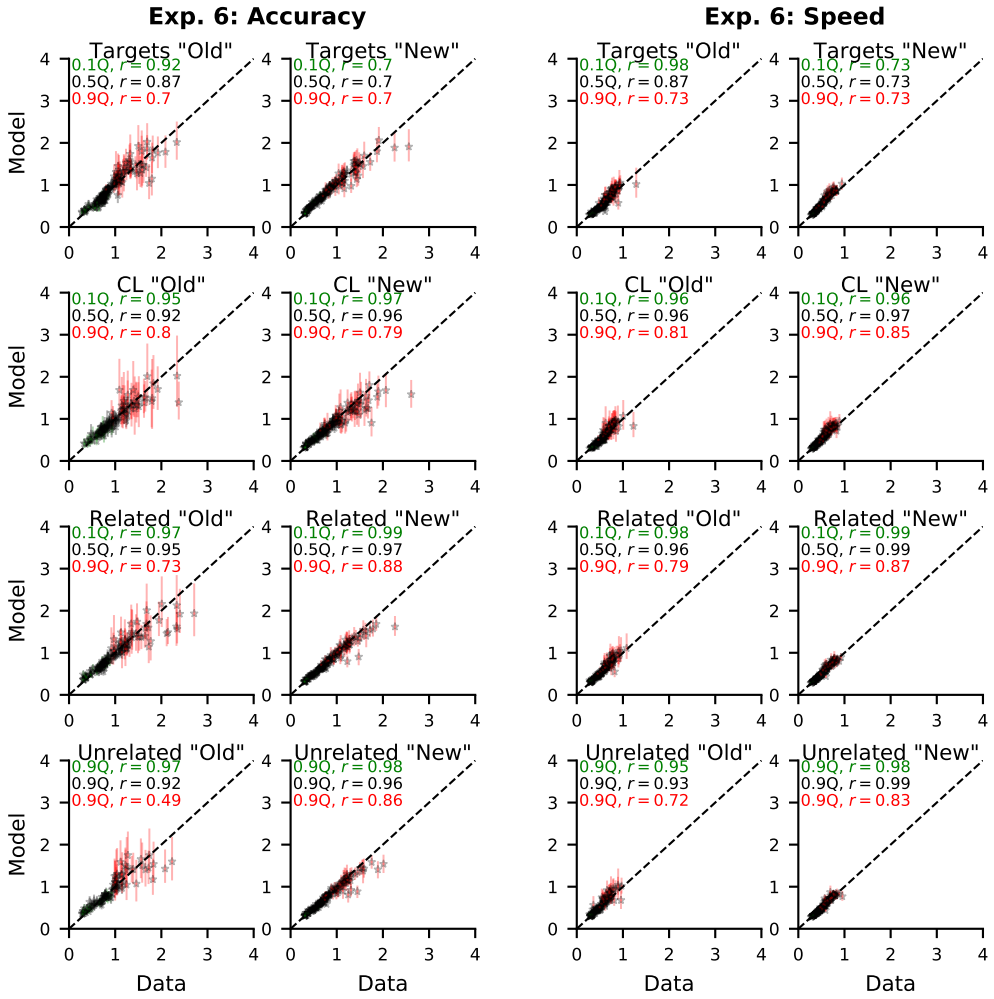

Figure 13. Individual participant RT distributions (summarized using the .1, .5, and .9 quantile) for each trial type from both the data and the model from Experiment 6. These results were collapsed across both the semantic and structural DRM tasks. Note that CL = critical lure, rel. = related lure, unrel. = unrelated lure, and unrel. CL = unrelated critical lure. Error bars represent 95% highest density intervals (HDIs).

Cross-Task Correlations

The cross-task correlations for true and false recognition from both the data and the model can be seen in Figure 14. Interestingly, there was a single case that resembled the data of Ballou and Sommers (2008), namely the accuracy-emphasis condition, which showed reasonable correlations for true recognition ($r = .7$) and a very low correlation for false recognition ($r = .12$). However, this pattern was wildly different in the speed-emphasis

condition, where the correlation for false recognition was considerably higher ($r = .54$).

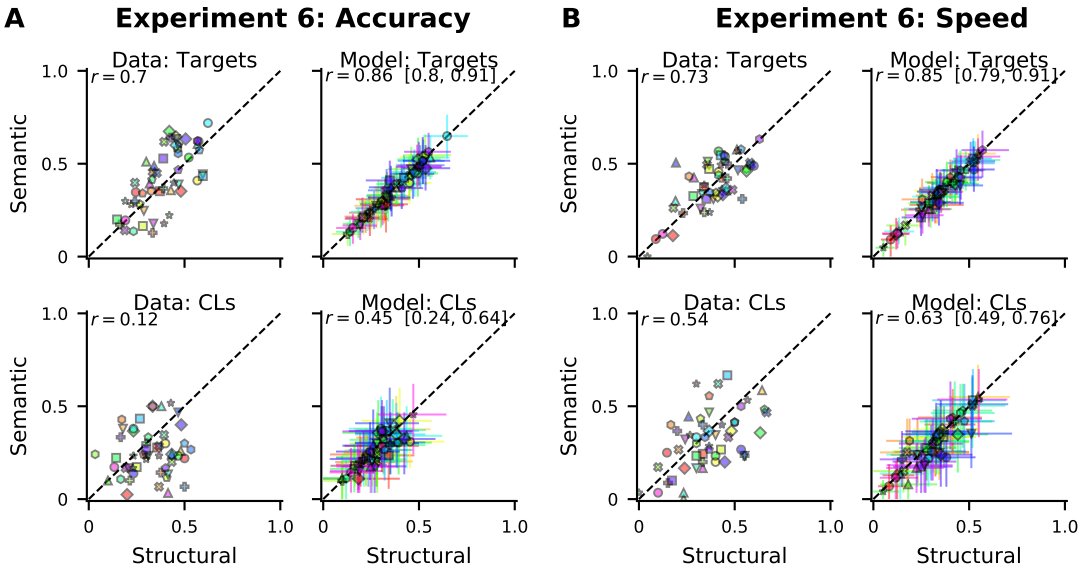

Figure 14. Scatterplots showing the relationship between the semantic (x-axis) and structural (y-axis) DRM tasks for the speed- and accuracy-emphasis conditions in Experiment 6. For each task, we compared the hit rate to target items (top) and the false alarm rate to critical lures from studied categories (bottom). For the model (which is the augmented model from Experiment 5), we reported both the mean and 95% highest density interval (HDI) of the correlation coefficient. Error bars show the 95% HDIs on the model's hit and false alarm rate predictions.

The model provides a reasonable account of the cross-task correlations in the speed-emphasis condition. While the mean correlation estimate for false recognition is higher than the data, the 95% HDI is extremely wide, and notably includes the correlation estimate of the data. The model also predicts stronger correlations under speed-emphasis than accuracy-emphasis, but the model notably predicts considerably higher cross-task correlations for false recognition in the accuracy-emphasis condition (mean $r = .43$). While the 95% HDI is wide, it does not include the data's correlation estimate.

We additionally pursued another variant of the augmented model where we allowed parameters related to memory encoding varied across the semantic and structural DRM tasks. This model revealed impaired memory strength in the structural lists (mean $m^{\mu}_{\text{structural}} = .676$, 95% HDI = [.622, .73]). In addition, it revealed lower precision in the

structural DRM task than in the semantic DRM task. This was observed in both the speed-emphasis (mean $\tau^{\mu}_{\text{structural}} = 1.13$, 95% HDI = [.978, 1.301], mean $\tau_{\text{semantic}} = 1.33$, 95% HDI = [1.16, 1.48]) and accuray-emphasis condition (mean $\tau^{\mu}_{\text{structural}} = 1.59$, 95% HDI = [1.47, 1.72], mean $\tau_{\text{semantic}} = 1.84$, 95% HDI = [1.70, 1.98]).

The cross-task correlations from the model can be seen in Figure 15. The cross-task correlations in this model are lower and more closely resemble the data. In the accuracy-emphasis condition, the cross-task correlation for true recognition ($r = .7$) is within the 95% HDI of the model (.65-.84). While the cross-task correlations for false recognition are lower than before, the correlation from the data ($r = .12$) is only slightly below the 95% HDI from the model (.13-.58).

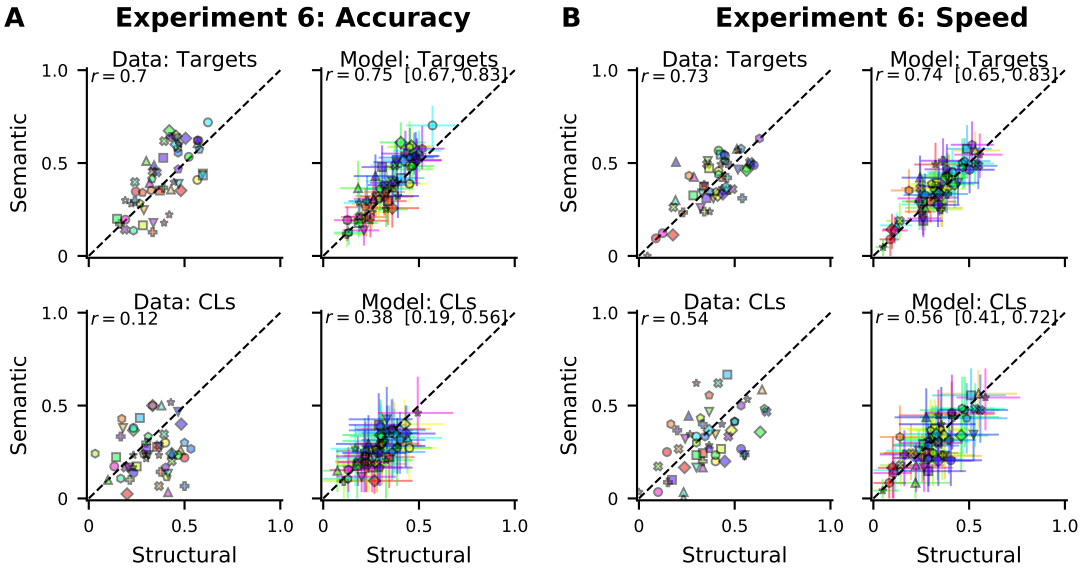

Figure 15. Scatterplots showing the relationship between the semantic (x-axis) and structural (y-axis) DRM tasks for the speed- and accuracy-emphasis conditions in Experiment 6. For each task, we compared the hit rate to target items (top) and the false alarm rate to critical lures from studied categories (bottom). For the model (which is the augmented model from Experiment 5 but with allowance for impaired encoding in the structural lists), we reported both the mean and 95% highest density interval (HDI) of the correlation coefficient. Error bars show the 95% HDIs on the model's hit and false alarm rate predictions.

We would again like to emphasize that these cross-task correlations have varied

widely across both experiments and conditions. Other factors – such as differences in stimulus materials as well as differences in encoding across the lists – may be responsible for low cross-task correlations for false recognition. Proponents of the view that semantic and structural false recognition have the burden of explaining why cross-task correlations vary considerably and why they can be strongly positive in some cases.

G: Parameter Recovery

We performed a parameter recovery to evaluate whether our model's parameters can be reasonably identified. Strong parameter recovery also indicates that a model can be used as a measurement model in future applications. The basic principle of parameter recovery is to generate synthetic datasets from the model's estimated parameters, fit those synthetic datasets with the same model and methodology used in the main text, and then evaluate whether the estimated parameters resemble the parameters that were used to generate the synthetic datasets. If the estimated parameters show a strong relationship to the original generating parameters, then it can be concluded that the model is well identified.

A complication with parameter recovery in a Bayesian context is that an entire posterior distribution of parameter values is estimated. It is generally not feasible to perform parameter recovery on every posterior sample as this could involve thousands of parameter recovery exercises. We simplified matters considerably by restricting our choice of parameters to a single set. Specifically, for each participant, we used the parameter vector with the highest likelihood and subsequently generated a synthetic dataset that is of the same size as the original data. These synthetic data were then fit in the same manner as the real data were fit in the main text.

Parameter recovery results for Experiments 1, 2, and 3 can be seen in Figures 16, 17, and 18, respectively. The results are presented on scatterplots where the generating values used to produce the synthetic datasets are on the x-axis, while the estimated values from the recovery exercise are on the y-axis. For each parameter, we plotted the 95% highest

density interval (HDI) estimated from the recovery exercise. The overall performance of the recovery exercise was assessed by calculating the correlation between the mean posterior estimate and the generating parameter value.

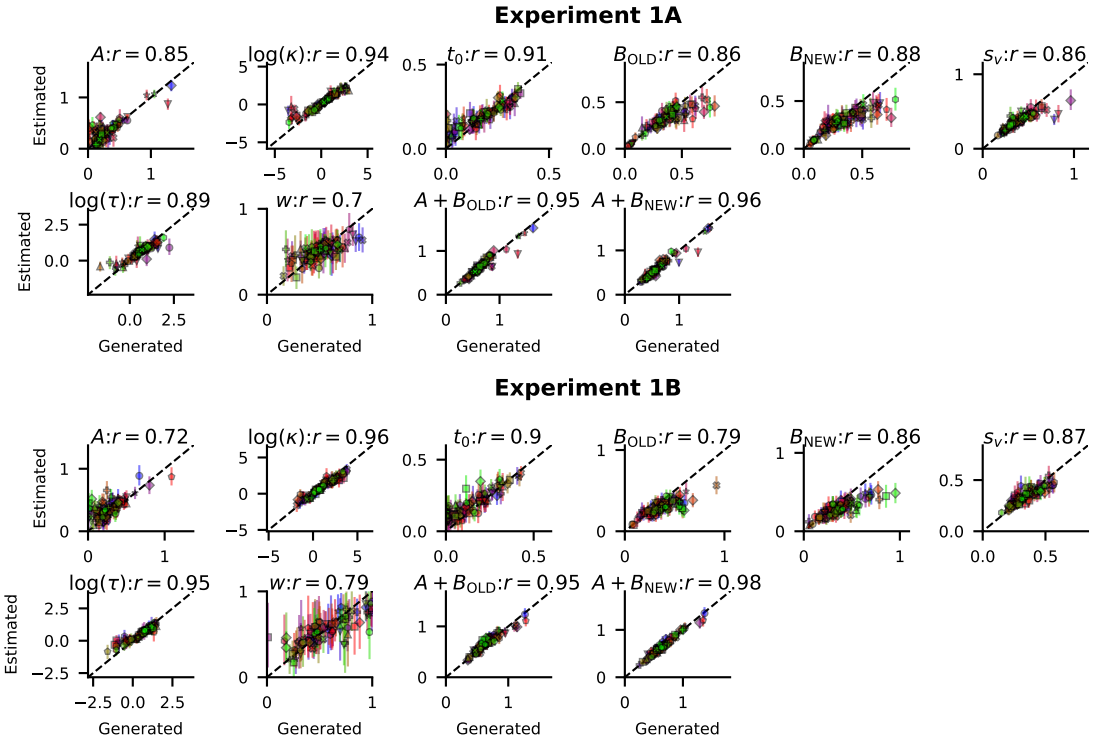

Figure 16. Parameter recovery for Experiment 1, which used the semantic DRM task. The generating values can be found on the x-axis while the estimated values can be found on the y-axis. Error bars depict the 95% highest density interval (HDI).

One can see that the model is well identified on the whole. Recovery of the τ and κ parameters was often excellent. However, there were some parameters that did not recover as well. We did note that when the generating values of B were higher, the estimates from the recovery exercise were often lower than the generating values. Likewise, some of the lower generating values of A were under-estimated in the recovery exercise. We strongly suspected that there was a tradeoff between these two parameters – low values of B can be compensated with high values of A. For this reason, we additionally plotted the sum of these two parameters ($A + B$) as a measure of the overall height of the response threshold.

We found that recovery was excellent for this sum ($r > .95$ in many cases).

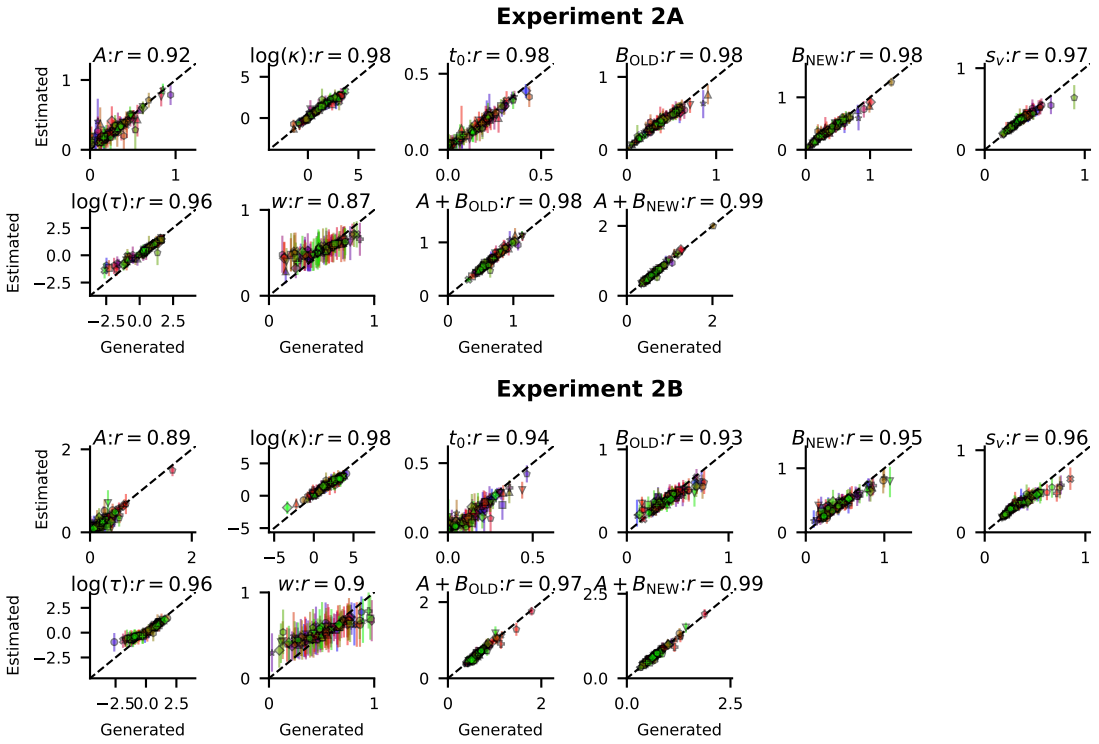

Figure 17. Parameter recovery for Experiment 2, which used the structural DRM task. The generating values can be found on the x-axis while the estimated values can be found on the y-axis. Error bars depict the 95% highest density interval (HDI).

The w parameter in Experiment 1 and 2 had much higher estimates of uncertainty (e.g., wider 95% HDI intervals) than in Experiment 3. As mentioned in the main text, a high value of w implies strong semantic false recognition, while a low value of w implies strong structural false recognition. Experiment 3 likely exerted much more constraint on the estimate of the w parameter because participants were tested on both DRM tasks. Experiments 1 and 2 only tested semantic or structural DRM tasks, so only one type of false recognition was strongly observed for each participant.

Parameter recovery for Experiments 4 and 5 can be seen in Figures 19 and 20, respectively. The parameter recovery exercise for Experiment 5 was limited to the augmented model, since that was the preferred model in the main text. One can see that

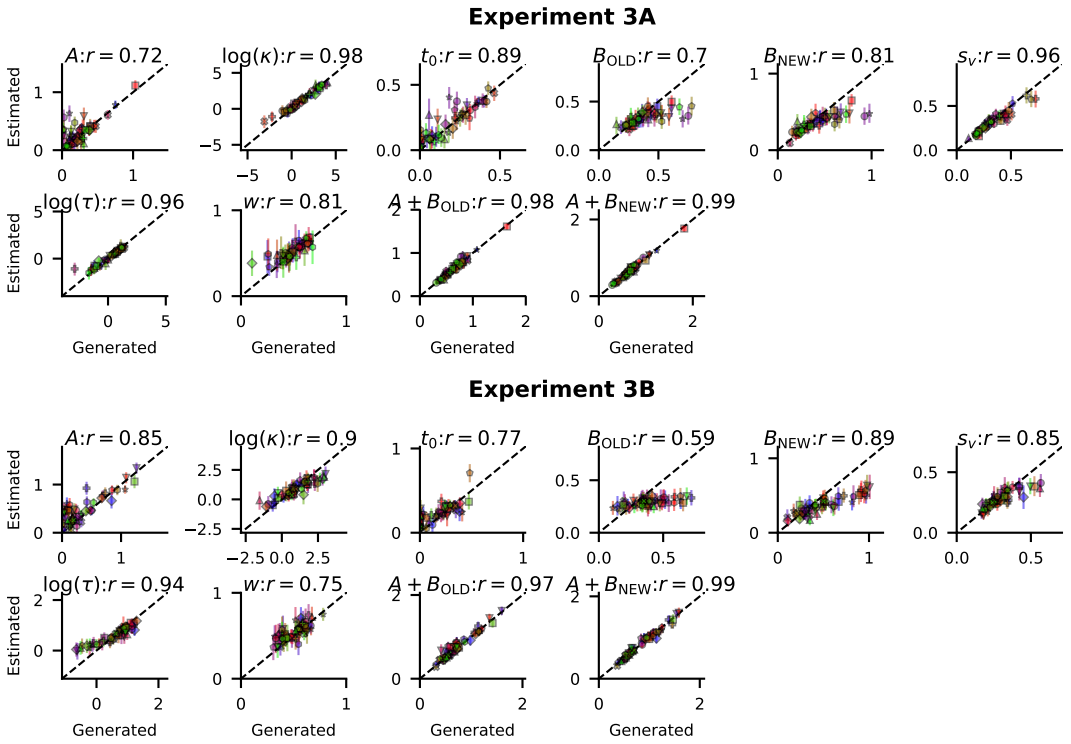

Figure 18. Parameter recovery for Experiment 3, which used both the semantic and structural DRM tasks. The generating values can be found on the x-axis while the estimated values can be found on the y-axis. Error bars depict the 95% highest density interval (HDI).

despite the fact that the augmented model had the largest number of parameters (14), it showed excellent recovery, with some parameters being almost perfectly recovered. This suggests that the data and experimental design provided strong constraints on the model's parameter estimates.

H: Changes in False Recognition Across Test Trials

In the main text and the General Discussion, we argued that the higher false alarm rates for unrelated critical lures (e.g., critical lures from unstudied categories) may arise from the storage of test items. That is, if test items are stored and added to the memory set, and an unrelated critical lure is tested after some of the exemplars from its own category have already been tested, then the storage of these traces may greatly increase the

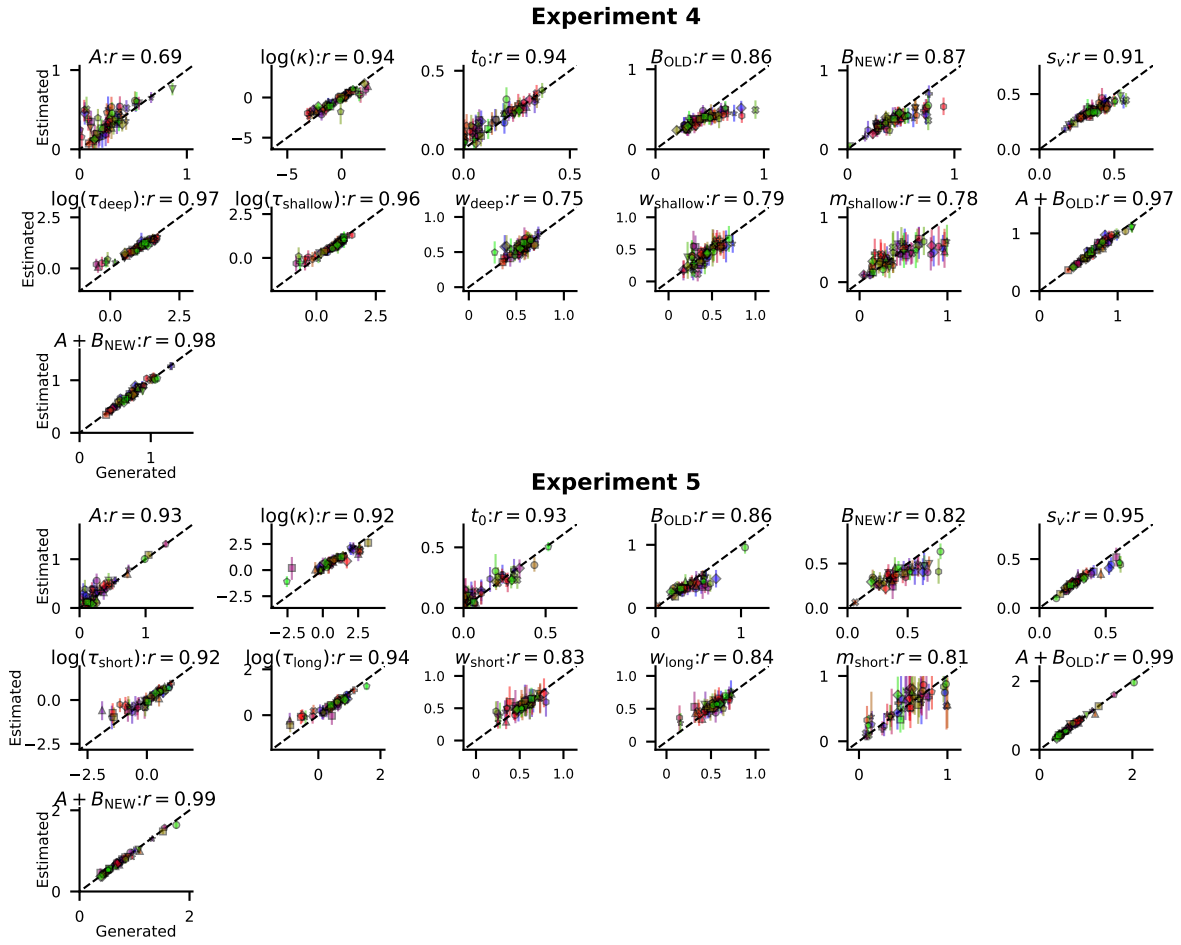

Figure 19. Parameter recovery for Experiments 4 and 5, which used both the semantic and structural DRM tasks and manipulated depth of processing (Experiment 4) or study time (Experiment 5). The generating values can be found on the x-axis while the estimated values can be found on the y-axis. Error bars depict the 95% highest density interval (HDI).

global matching signal for unrelated critical lures, leading to higher rates of false recognition.

If this account is correct, then unrelated critical lures should exhibit low rates of false recognition early in the test list when they are less likely to have been tested after other exemplars from their own category have been stored. In the later parts of the test list, however, we should expect higher rates of false recognition for unrelated critical lures, as it is far more likely that they have been tested after other exemplars from their category have

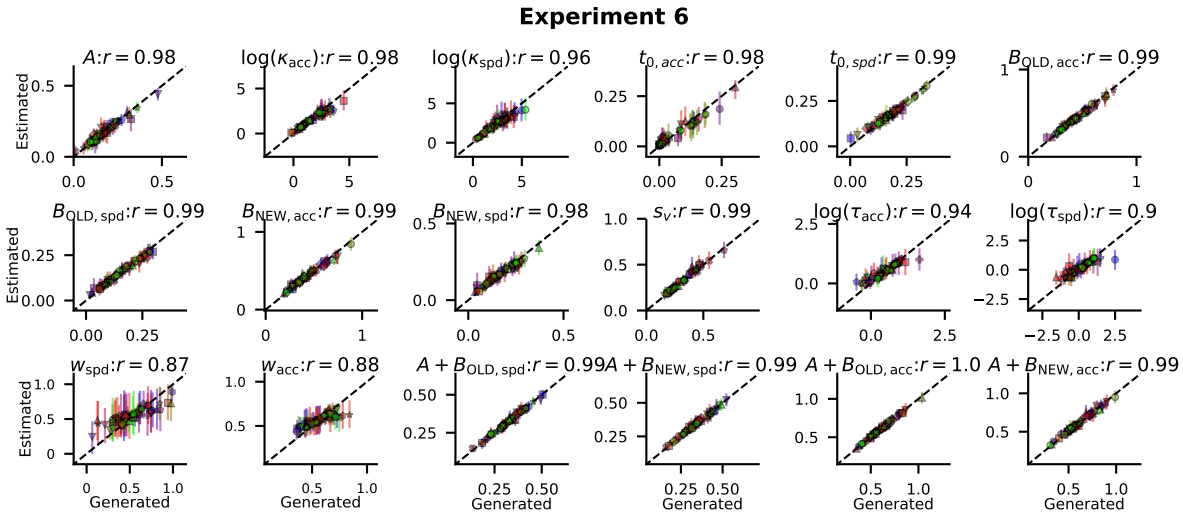

Figure 20. Parameter recovery for Experiment 6, which used both the semantic and structural DRM tasks and manipulated time pressure at test. The generating values can be found on the x-axis while the estimated values can be found on the y-axis. Error bars depict the 95% highest density interval (HDI).

already been tested.

To test for this possibility, we re-analyzed the false recognition rates from our experiments for different regions of the test list. We did not consider the data from Experiment 3B as the test lists only included 27 test trials, which was not long enough to observe substantial changes over the testing period. For the remaining datasets, we divided the test list into thirds because it was the easiest way to divide our test lists – Experiments 1A and 2A have 117 trials in the test list (which divide into three blocks of 39 trials) while all other remaining datasets have 63 trials in the test list (which divide into three blocks of 21 trials).

False recognition for critical lures, related lures, unrelated lures, and unrelated critical lures for each dataset can be seen in Figure 21, with the semantic DRM task in the left column and the structural DRM task in the right column. A number of trends are evident. First, false recognition of critical lures and related lures is mostly stable. If any changes are evident, it's that false recognition of critical lures *decreases* in some cases, such

as in Experiment 2A.

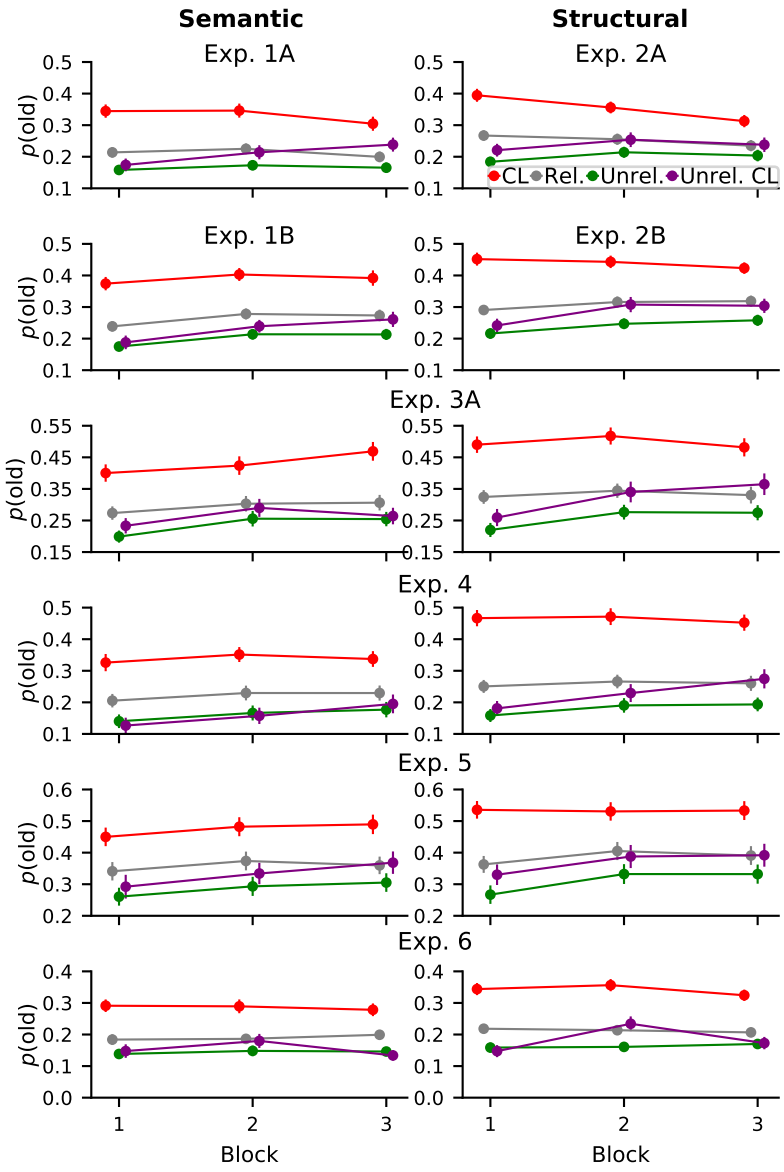

Figure 21. False recognition rates for each third of the test list for the semantic DRM task (left) and the structural DRM task (right) for Experiments 1, 2, 3A, 4, 5, and 6. Error bars represent the standard error of the mean.

Of greater interest – however – is the pattern for unrelated critical lures. In the majority of the cases, false recognition of unrelated critical lures increases to the greatest extent of each of the lure types. The only exception is Experiment 6 – where time pressure was manipulated – which does not show many changes throughout the test list for any of

the trial types. However, for the majority of the other cases the false recognition of the unrelated critical lures increases to a greater extent than to the unrelated lures.

I: Details of Huff and Ashcenbrenner Dataset

For the data of Huff and Aschenbrenner (2018), while the main methodology is described in the main text, there were a few other miscellaneous issues involving the application of our model to these data that are described here.

First, the word "cave" was found to be tested twice as both a "target" and a "critical lure" on every single test. While we had considered re-classifying the second occurrence of this test item, we found it was simpler to just remove all trials that contained the word "cave."

Second, the word "chest" occurred in two different categories. In every case where it was tested as an "unrelated lure", it had actually occurred on the study list because of its occurrence in another category. For this reason, we re-classified all test trials of "chest" as target trials.

Third, there were some complications in constructing semantic representations of the words. Unlike our experiments, a number of items were actually two words – "blue jay", "coffee table", "egg beater", "frying pan", "onion salt", "rolling pin", "bay leaves", "can opener", "end table", "lima beans", "measuring cup", "terry cloth", and "short story." For several of these words, we were able to overcome this problem by removing the space and obtaining a semantic representation of the compound word (e.g., "bluejay", "coffeetable"). However, there were a few words where the compound words were not present in the Common Crawl corpus of the fasttext library – these include "onion salt", "rolling pin", "lima beans", and "measuring cup." For these words, we obtained the semantic vectors of the individual words and summed them together.

Additional Exclusions

While the main text details the performance-related exclusions, we found that there were some aspects of the data that rendered it unusable for some participants. For participant 34 of the younger adults dataset, around 18 words were described as "unrelated lures" when in fact they were studied items. Given that this was a large number of items, we decided to simply reject the participant entirely.

J: List Blocking Experiment

Prior to running Experiment 1 in the main text, we ran an experiment where we manipulated "list blocking" between subjects in the semantic DRM paradigm. In the blocked condition, all items from each category were presented consecutively. In the spaced condition, the items were randomized throughout the study list.

Many aspects of the design and method were virtually identical to our Experiment 1, with the following exceptions. Participants studied eight exemplars per category from a total of four categories, producing study lists containing 64 items. A different set of categories were employed from the main text – a total of 60 categories were used that each contained sixteen items. Participants studied a total of 32 categories, leading to 32 critical lure trials as opposed to 60 trials like in the main text.

The data are publicly available on our OSF repository (`https://osf.io/rek4y/`).

Method

Participants. A total of 183 participants (81 in the blocked condition, 102 in the spaced condition) completed the experiment in exchange for course credit. Similar to the main text, we rejected participants for poor performance ($d' < .1$). This led to the rejection of seven participants in the blocked condition and eight participants in the spaced condition, leading to a final sample of 168 participants (74 in the blocked condition, 94 in the spaced condition).

Materials. 60 categories containing 16 exemplars were selected from the De Deyne, Navarro, Perfors, Brysbaert, and Storms (2019) free association norms.

Procedure. Similar to Experiments 1-3 in the main text, participants studied words for 800 ms with a 200 ms inter-stimulus interval (ISI). The study list contained four categories of eight exemplars each. The exemplars were drawn randomly from each category. In the blocked condition, all items were presented from a given category before any items from the other categories were presented, but the presentation order within a category was randomized for each study list. In the spaced condition, all items were randomized across the study list.

On the test list, participants were tested on all eight targets from each studied category, the remaining eight items from each studied category were tested as related lures, and sixteen exemplars from two unstudied categories were tested as unrelated lures. The critical lures from each studied category were tested along with the critical lures from the unstudied categories. Thus, on each test list, participants were tested on 32 targets, 32 related lures, 32 unrelated lures, 4 critical lures from studied categories, and 2 critical lures from unstudied categories.

Participants completed a total of eight study-test cycles, leading to a total of 256 target trials, 256 related lure trials, 256 unrelated lure trials, 32 critical lure trials, and 16 unrelated critical lure trials.

Results

Group-averaged hit and false alarm rates can be seen in Figure 22. What is immediately obvious is that the list blocking effect on the false recognition of critical lures is almost entirely absent – false recognition of critical lures was only slightly higher ($M = .39$) in the blocked condition relative to the spaced condition ($M = .38$). A Bayesian independent samples t-test found evidence for a null effect of blocking on false recognition ($BF_{10} = .286$), indicating the null hypothesis is 3.5 times more likely than the alternative

hypothesis.

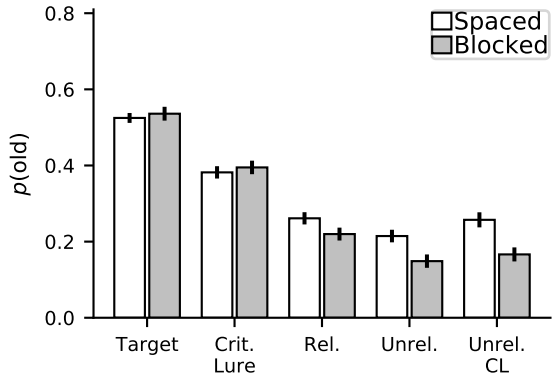

Figure 22. Group-averaged hit and false alarm rates from our supplementary experiment. List blocking (blocked vs. spaced) was manipulated between subjects. Error bars depict the standard error of the mean.

The effects of list blocking on true recognition were similarly slight – hit rates exhibited a slight advantage in the blocked condition ($M = .54$) relative to the spaced condition ($M = .52$), although a Bayesian t-test similarly indicated support for a null effect on hit rates ($BF_{10} = .227$).

Nonetheless, there were effects of list blocking on the rejection of unrelated lures. False recognition of unrelated lures was lower in the blocked condition ($M = .15$) relative to the spaced condition ($M = .15$). A Bayesian independent samples t-test indicated that this is a true effect ($BF10 = 33.71$).

Why is there is a discrepancy from prior results? Both Mather, Henkel, and Johnson (1997) and Tussing and Greene (1997) found substantial effects of list blocking on false recognition of critical lures. There are a number of important differences between our design and these designs that were alluded to in the main text.

The first difference is the fact that in the aforementioned studies, the blocked design always presented the same studied items, namely the strongest associates to each critical lure. Moreover, they were always presented in the same order, typically from the strongest to the weakest associate. We argue in the main text that such a procedure makes it more

likely that participants will think of the critical lure while studying the list, which can lead to encoding-based false recognition errors during a later test list.

The second difference is the fact that we used faster presentation times. More rehearsal is possible with longer presentation times, and when more rehearsal is afforded, it makes it more likely that participants will erroneously rehearse the critical lure during the study phase (Marsh & Bower, 2004; Seamon et al., 2002). A related point is the fact that the two above studies had breaks between each category in the blocked condition, which allows for more rehearsal time and also clearly indicates to participants that there are differences between each category.

The final difference is the fact that we have more data. We tested 32 critical lure trials per participant as opposed to 10 critical lure trials per participant in the other experiments. We also had large-ish samples of participants.

We do not mean to suggest that this experiment is the "last word" on list blocking. As we mention in the main text, we argue that future work is necessary that manipulates some of the variables that may influence list blocking effects, such as study time, presentation order within each category, as well as breaks between the categories.

K: Signal Detection Theory Model of Participant- and Item-Level Effects

In the main text, we tested our memory model against variation in critical lures and obtained superficially low R^2 values. However, it was unclear whether these low values can be attributed to poor performance of the model or whether the data are sufficiently noisy to prevent any model from accounting for such variation. The latter possibility refers to the concept of a "noise ceiling" – where noise in the data prevents a model from achieving perfect performance.

We attempted to measure the noise ceiling by developing a model that can account for both participant- and item-level variation in the data. Specifically, we adopted a signal detection theory (SDT) approach. In SDT models, each class of items is associated with a

latent strength distribution (typically a Gaussian distribution) that represents across-trial variability in memory strength. On each trial, the memory strength is compared to a decision criterion and an "old" response is produced if the strength exceeds the criterion but a "new" response is produced otherwise.

In our SDT model, conditions – and in some cases, items – all receive their own corresponding Gaussian distribution. This Gaussian distribution has a standard deviation of 1 and a location that is determined by the parameters of the model. An exception is the unrelated lures, which we fixed to a mean of 0 in order to identify the model.

For targets, related lures, and unrelated critical lures, we use the parameter d_{target}, d_{related} and d_{ucl} to denote the means of their underlying distributions. We did not estimate any individual item variation for each of these item classes, but they did vary by participant. In addition, we estimated a criterion C for each participant.

For critical lure items, we estimated both a d parameter d_{cl} for each and for each critical lure item j. Thus, the Gaussian distribution $d^{ij} = d^i + d^j$.

We jointly estimated all of these parameters in a hierarchical Bayesian framework using DE-MCMC, which is the same as the text. Specifically, each participant-level parameter was sampled from group-level parameters with mean μ and standard deviation σ:

$$d^i_{\text{target}} \sim \text{Normal}(d^\mu_{\text{target}}, d^\sigma_{\text{target}}) \tag{1}$$

$$d^i_{\text{related}} \sim \text{Normal}(d^\mu_{\text{related}}, d^\sigma_{\text{related}}) \tag{2}$$

$$d^i_{\text{cl}} \sim \text{Normal}(d^\mu_{\text{cl}}, d^\sigma_{\text{cl}}) \tag{3}$$

$$d^i_{\text{ucl}} \sim \text{Normal}(d^\mu_{\text{ucl}}, d^\sigma_{\text{ucl}}) \tag{4}$$

$$C^i \sim \text{Normal}(C^\mu, C^\sigma) \tag{5}$$

Each group-level distribution for the participants had the following near

non-informative prior distribution:

$$C^{\mu} \sim \text{Normal}(.5, 2.0) \tag{6}$$

$$d^{\mu}_{\text{target}} \sim \text{Normal}(1.5, 2.0) \tag{7}$$

$$d^{\mu}_{\text{related}} \sim \text{Normal}(1.5, 2.0) \tag{8}$$

$$d^{\mu}_{\text{ucl}} \sim \text{Normal}(0, 1.5) \tag{9}$$

$$d^{\mu}_{\text{cl}} \sim \text{Normal}(1.0, 2.0) \tag{10}$$

$$d^{\sigma}, C^{\sigma} \sim \text{Gamma}(1.0, 1.0) \tag{11}$$

Each item-level parameter was sampled from its own respective group-level distribution. In other words, there were a total of 90 d parameters estimated – one for each critical lure word – which all came from a separate group-level distribution with a mean that is fixed at zero but with a freely estimated standard deviation ς.

$$d^{j}_{\text{cl}} \sim \text{Normal}(0, d^{\varsigma}_{\text{cl}}) \tag{12}$$

where the item group-level distribution ς was sampled from the following prior distribution:

$$d^{\varsigma}_{\text{cl}} \sim \text{Gamma}(1.0, 1.0) \tag{13}$$

Fixing the mean of the item group-level distribution to zero is critical for identifying the model. Accordingly, the d^{j}_{cl} parameters can be interpreted as differences from the participant parameters d^{i}_{cl}.

When running DE-MCMC, a total of 30 chains were run with 1,500 burn-in iterations. After the burn-in phase, a total of 1 in every 10 samples was collected until 750 samples were collected from each chain. A model was considered converged if its

Gelman-Rubin (G-R) diagnostic was below 1.2 for all participant and group-level parameters. This criterion was satisfied for all models, with the G-R statistic close to 1.0 in many cases for both participants and items.

References

Ballou, M. R., & Sommers, M. S. (2008). Similar phenomena, different mechanisms: Semantic and phonological false memories are produced by independent mechanisms. *Memory & Cognition*, *36*, 1450-1459.

De Deyne, S., Navarro, D. J., Perfors, A., Brysbaert, M., & Storms, G. (2019). The Small World of Words english word association norms for over 12,000 cue words. *Behavior Research Methods*, 987-1006.

Elhalal, A., Davelaar, E. J., & Usher, M. (2014). The role of the frontal cortex in memory: an investigation of the Von Restorff effect. *Frontiers in Human Neuroscience*, *8*, 1–20.

Farrell, S., & Lewandowsky, S. (2002). An endogenous distributed model of ordering in serial recall. *Psychonomic Bulletin & Review*, *9*, 59–79.

Huff, M. J., & Aschenbrenner, A. J. (2018). Item-specific processing reduces false recognition in older and younger adults: Separating encoding and retrieval using signal detection and the diffusion model. *Memory & Cognition*, 1287-1301.

Marsh, E. J., & Bower, G. H. (2004). The role of rehearsal and generation in false memory creation. *Memory*, *12*, 748–761.

Mather, M., Henkel, L. A., & Johnson, M. K. (1997). Evaluating characteristics of false memories: Remember/know judgments and memory characteristics questionnaire compared. *Memory & Cognition*, *25*(6), 826–837.

Seamon, J. G., Lee, I. A., Toner, S. K., Wheeler, R. H., Goodkind, M. S., & Birch, A. D. (2002). Thinking of critical words during study is unnecessary for false memory in the DRM procedure. *Psychological Science*, *13*, 526–531.

Tussing, A. A., & Greene, R. L. (1997). False recognition of associates: How robust is the effect? *Psychonomic Bulletin & Review*, *4*, 572-576.