Supplemental Material

Additional details about sample size determination and our participant pool

We determined our sample size using a power analysis based on the findings from previous research with the A-A' test format comparing 3 vs. 1 repetitions (Loiotile & Courtney, 2015). Here, we used the function `pwr.t.test` from the pwr package in R with power=0.9 and a significance level of .001, which revealed that we needed to have a minimum sample size of 10 participants. Given that we were additionally interested in testing several novel test formats here, we aimed to collect data from at least 50 participants, thus our sample of 66 participants is greater than our minimum a priori threshold.

The university's Institutional Review Board approved all procedures before data collection began. Participants were recruited from the Psychology Department's participant pool and received course credit for participation. We excluded 11 additional participants from the analysis due to technical errors (*n*=7), change in instructions (*n*=3), or failure to meet the inclusion criteria (*n*=1).

Predictions from Competitive Trace Theory

Here, we also compared the predictions of MINERVA 2 with the predictions from Competitive Trace Theory (CTT). Consistent with Prediction 1 from MINERVA 2, CTT would predict enhanced recognition of a target with multiple exposures when paired with a foil (i.e., better performance on A3x-X vs. A1x-X; Reagh & Yassa, 2014). CTT would also predict better performance when the original version of the non-corresponding lure (B) was presented with 1-repetition than 3-repetitions (Predictions 7-8; i.e., better performance on A1x-B'1x vs. A1xB'3x and A3x-B'1x vs. A3x-B'3x). However, contrary to MINERVA 2, CTT would predict poorer performance on A3x-A' vs. A1x-A' and A3x-B'3x vs. A1x-B'1x (Predictions 2-3) because 3-repetitions would lead to semantization, thus resulting in a loss of episodic detail and a higher likelihood of selecting the non-corresponding lure (Reagh & Yassa, 2014; Yassa & Reagh, 2013). Because CTT is a verbal theory (Reagh & Yassa, 2014; Yassa & Reagh, 2013), it was difficult to know exactly what the model would predict for the MINERVA

2 Predictions 4-6, thus again highlighting the importance of computational modeling for "open theorizing" and allowing the generation of testable predictions based on a given set of assumptions (Guest & Martin, 2021).

Methodological details for the simulations with MINERVA 2

We have previously shown that MINERVA 2 can account for a variety of findings with the Mnemonic Similarity Task, including the test-format effect (Huffman & Guan, 2025; Huffman & Stark, 2017), the effect of target-lure similarity (Huffman & Guan, 2025), and age-related differences observed during childhood (Rollins, Huffman, et al., 2023) and healthy aging (Huffman & Stark, 2017). Thus, we generated predictions of performance across the following test formats: A1x-X, A3x-X, A1x-A', A3x-A', A1x-B'1x, A3x-B'3x, A3x-B'1x, and A1xB'3x. Briefly, MINERVA 2 assumes that memories are stored as distributed representations of features, here, modeled as -1, 0, and +1 (in our case, we simulated equal probability of each of these 3 features for each element of the item vector and we used a vector length of 20 features, as in Huffman & Guan, 2025; Huffman & Stark, 2017). MINERVA 2 further assumes that the features are encoded with a given probability representing the learning rate, L (here, each feature was encoded with probability of L and not encoded [i.e., set to 0] with probability 1 - L), which we set to 0.7. We note that in our previous work, we used a learning rate of 0.65 (Huffman & Guan, 2025; Huffman & Stark, 2017), but we increased L slightly here to accommodate both the longer list length and the variable list strength here. We generated similar lures as in past work by setting the probability of changing a feature parameter, δ, to 0.16 (Huffman & Guan, 2025; Huffman & Stark, 2017). MINERVA 2 is a multiple-trace model, which assumes that each experience adds a new trace or row to a memory vector (i.e., the columns represent the features and the rows represent the events). For our empirical study, we included 25 stimuli within each condition. Thus, here, we simulated 25 stimuli with one repetition and 25 stimuli with 3 repetitions. On the back end, we then

generated and arranged the test items to allow for all of our stated test formats (see above). We ran the model on 1,000 simulated participants.

For the testing phase, we used the standard equations from the original MINERVA 2 model (Hintzman, 1984, 1988). First, for each probe item in a simulated trial, we calculated the similarity of the probe to the memory matrix, s_i, using the following equation (Hintzman, 1984, 1988):

$$s_i = \frac{p \cdot T_i}{n_i} \quad (1)$$

where p is the probe item, T_i is one of the items of the memory matrix (i.e., row i of the memory matrix, T), and n_i is the number of nonzero features in either p or T_i (i.e., the similarity function is a normalized dot product). We then calculated the activation, a_i, for each of the items with the following equation (Hintzman, 1984, 1988):

$$a_i = (s_i)^3 \quad (2)$$

We then calculated the global match, g, or summed similarity of the probe item to the entire memory matrix using the following equation (Hintzman, 1984, 1988):

$$g = \sum_{i=1}^{M} a_i \quad (3)$$

where M indicates the number of items in the memory matrix (i.e., then number of rows). For each trial, we calculated the global memory match, g, for each of the two simulated probe items. We then calculated the performance using the following equation (cf. Hintzman, 1988; Huffman & Guan, 2025; Huffman & Stark, 2017; Rollins, Huffman, et al., 2023):

$$Pr\{A > X\} = \frac{1}{M}\sum_{i=1}^{M}\left[I\left(g_{A_i} > g_{X_i}\right) + \left(0.5 \times I\left(g_{A_i} = g_{X_i}\right)\right)\right] \quad (4)$$

where M is the number of items in the memory matrix, $I()$ is the indicator function that outputs 1 if the condition is true and zero otherwise, g_{Ai} is the global match in response to the target item on trial i, and g_{Xi} is the global match in response to the distractor item (either an unrelated foil or a similar lure item) for trial i. Note that the left side of the equation calculates the sum of trials in which the

global match, *g*, is greater for the target item than the lure item and the right side of the equation simulates random guessing for trials in which the global match is the same for the target item and the lure item (i.e., on average, 50% of these would be correct, hence the 0.5 term here). For our follow up analysis in which we analyzed the data within our binomial mixed model framework, we instead randomly selected the correct vs. incorrect item on a trial-by-trial basis (because the binomial model requires an exact possible proportion of correct trials for each condition).

Follow-up analyses on the test format effect in each repetition condition

As we reported in the main manuscript, we observed a significant effect of the repetitions contrast (β = 0.74, $\chi^2(1)$ = 113.70, p < .001), the test-format contrast (β = 0.91, $\chi^2(1)$ = 102.15, p < .001), and a repetitions-contrast × test-format-contrast interaction (β = 0.20, $\chi^2(1)$ = 5.67, p = .017). To further explore the test format effects, we conducted follow-up analyses to test whether the test-format effect was significant within each repetition condition. We defined our models as follows:

```
model <- mixed(ProportionCorrect ~ TestFormat + (1 | ParticipantID),
  data=data, method='LRT', family=binomial, weight=data$n)
```

where `ProportionCorrect` is the proportion correct for a given format, `TestFormat` is a contrast of expected performance based on MINERVA 2 (+1 for A-X, 0 for A-A', -1 for A-B'), `(1 | ParticipantID)` is the random effect of participant (here we used random intercepts because more complex models failed to converge), and `weight=data$n` is a vector of the number of trials within each condition (here, we had 25 trials for each of these conditions). We found a significant effect of the test-format contrast within the 1-repetition data (β = .71, $\chi^2(1)$ = 204.46, p < .001). Likewise, the planned comparisons revealed that all 3 test formats were significantly different than each other within the 1-repetition data: A1x-X > A1x-A' (β = 1.10, $\chi^2(1)$ = 98.79, p < .001), A1x-X > A1x-B1x' (β = 1.54, $\chi^2(1)$ = 216.75, p < .001), and A1x-A' > A1x-B1x' (β = 0.43, $\chi^2(1)$ = 24.20, p < .001).

Altogether, these results support the predictions of MINERVA 2 that there will be a A-X > A-A' > A-B' test format effect within the 1-repetition condition (compare these results with Figure 2A). Next, we tested whether the test format-effect was also significant within the 3-repetition condition by running a similar analysis on the data from the 3-repetition condition. We found a significant effect of the test-format contrast for the 3-repetition condition (β = 0.91, $\chi^2(1)$ = 203.23, p < .001). Likewise, the planned comparisons revealed that all 3 test formats were significantly different than each other within the 3-repetition data: A3x-X > A3x-A' (β = 1.12, $\chi^2(1)$ = 63.87, p < .001), A3x-X > A3x-B3x' (β = 1.98, $\chi^2(1)$ = 211.65, p < .001), and A3x-A' > A3x-B'3x (β = 0.69, $\chi^2(1)$ = 42.16, p < .001). Altogether, these results support the predictions of MINERVA 2 that there will be a A-X > A-A' > A-B' test format effect within the 3-repetition condition (compare these results with Figure 2A). In summary, we found evidence to support the A-X > A-A' > A-B' test format effect in both the 1-repetition and 3-repetition condition, thus building on previous behavioral and modeling findings with the typical 1-repetition version of the MST (Huffman & Guan, 2025; Huffman & Stark, 2017; Rollins et al., 2019, 2024; but see Migo et al., 2009).

Analysis of the data from the MINERVA 2 simulations

Following the somewhat unexpected finding of the interaction in the empirical data (see main text), we went back to our original a priori modeling results and decided to analyze the data from the model's predictions (i.e., we had previously only visually interpreted the data to note the prediction of the test format effect and the repetition effect) using the original modeling parameters. We used the same mixed-effects modeling approach for the simulated data (except we only included random intercepts for participants here, since more complex random effects models failed to converge). Here, we found that MINERVA 2 made qualitatively similar predictions to the empirical results: we observed a significant effect of the repetitions contrast (β = 1.44, $\chi^2(1)$ = 6,044.28, p < .0001) and the test-format contrast (β = 1.04, $\chi^2(1)$ = 9,354.52, p < .0001). Moreover, we observed a significant

repetitions-contrast × test-format interaction (β = .80, $\chi^2(1)$ = 1,148.13, p < .0001). Furthermore, consistent with the empirical data, follow-up pairwise comparisons of estimated marginal means (using emmeans within R; here we used a categorical version of the Repetitions and Test Format for easier interpretation of the effects) revealed that the repetition effect was (numerically) largest for the A-X format (odds ratio = 59.88, p < .0001) followed by the A-A' format (odds ratio = 2.96, p < .0001) and smallest for the A-B' format (odds ratio = 59.88, p < .0001). Likewise, as in the empirical data, these findings are further supported by the positive β for the interaction term, thus suggesting the prediction of an interacting benefit of increased repetitions with the increasing performance across the test format contrast. We note that the overall magnitude of effects are larger for the simulations, but we included 1,000 simulated participants and we would like to emphasize that we were concerned with qualitative effects (i.e., rather than attempting to quantitatively fit the data).

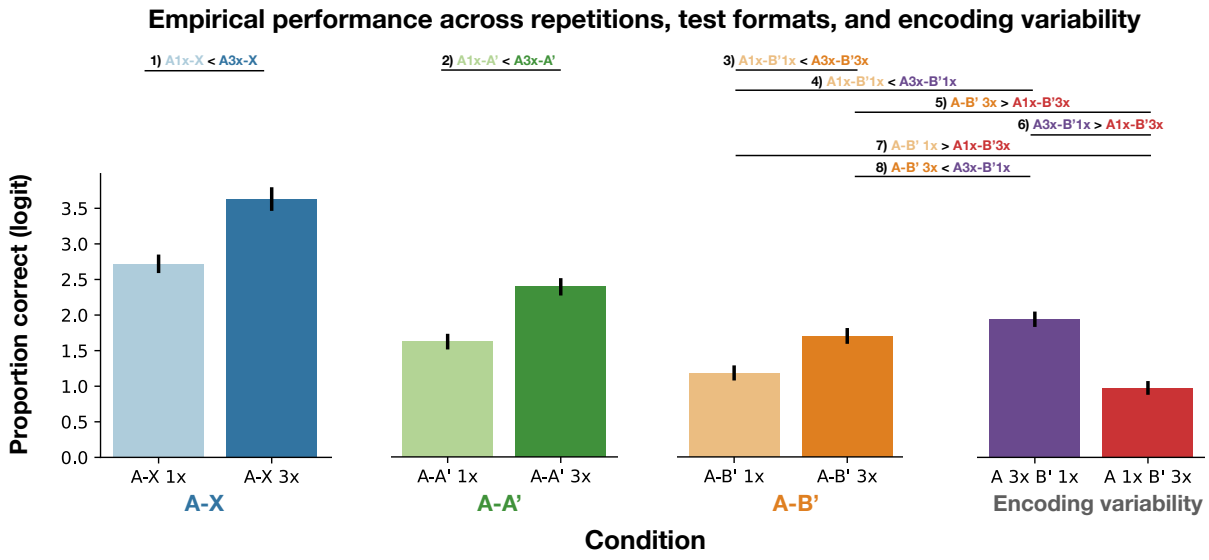
Lure Bins

We next tested whether performance varied as a function of target-lure similarity across all of the test formats with similar lures. Here, we used the "lure bins" based on behavioral performance from Stark et al. (2013). Specifically, Stark et al. (2013) created the "lure bins" from the most similar in lure bin 1 (highest "old" responses) to the most dissimilar in lure bin 5 (the lowest "old" responses). Here, we set up a contrast for lure bins 1 to 5: [-2, -1, 0, 1, 2] (the `LureBinContrast` variable in the equation below), which allowed us to test our specific prediction that performance would be lowest for lure bin 1 and progressively better through lure bin 5. We again ran our analysis within afex with a binomial family in R, using the following model:

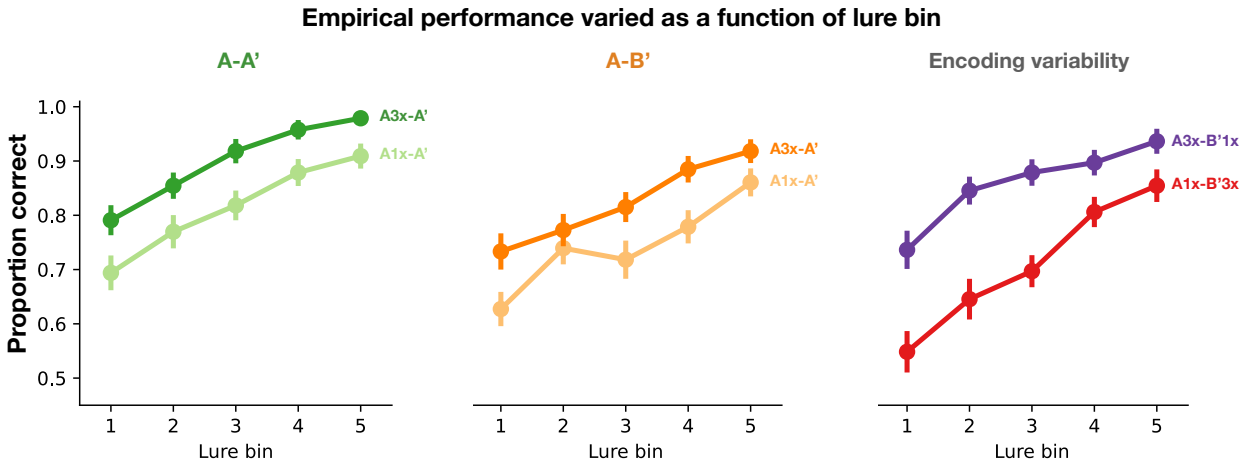
```
model <- mixed(ProportionCorrect ~ LureBinContrast + (1 | ParticipantID),
  data=data, method='LRT', family=binomial, weight=data$n)
```

Here, we had 5 trials within each lure bin for each test type. We ran this model separately within each condition and repetition to determine whether there was a significant effect of lure bin, which

would be consistent with MINERVA 2 (Huffman & Guan, 2025). We found a significant effect of the lure bin contrast in all of the conditions: A-A' 1x (β = 0.39, $\chi^2(1)$ = 68.19, p < .001), A-A' 3x (β = 0.64, $\chi^2(1)$ = 94.14, p < .001), A-B' 1x (β = 0.29, $\chi^2(1)$ = 47.44, p < .001), A-B' 3x (β = 0.38, $\chi^2(1)$ = 58.99, p < .001), A3x-B'1x (β = 0.62, $\chi^2(1)$ = 63.51, p < .001), and A1x-B'3x (β = 0.44, $\chi^2(1)$ = 107.97, p < .001; note: the original model failed to converge, so we dropped the lure bin 3 condition, in which case the model converged, thus our model was slightly different here, but included the contrast: [-1.5, 0.5, 0.5, 1.5]). Thus, the effect of lure bin is consistent across all 6 test formats with similar lures. Moreover, our results support recent work with several global matching models, including MINERVA 2, which all predicted that performance on test formats with targets and similar lures would vary as a function of the target-lure similarity (Huffman & Guan, 2025).



Supplemental Figure 1: Here, we replotted the data from Figure 1B on the logit scale (i.e., the original scale as our statistical analysis with the mixed-effects binomial models). The repetition-contrast × test-format-contrast format interaction effect for the A-X, A-A', and A-B' test format is clearer on this scale: the repetition effect was (numerically) largest for the A-X test format, followed by the A-A' test format, and smallest for the A-B' test format. The data depict the estimated marginal means and the 95% standard errors from our binomial models.



Supplemental Figure 2: Performance varied as a function of the lure bin across all 6 of the force-choice conditions with similar lures (i.e., a significant positive effect of the lure-bin contrast of [-2, -1, 0, +1, +2] across lure bins 1-5, which are ordered from most similar to least similar, based on previous research [see above]; all *ps* < .001). The data for the plots indicate the raw means and 95% standard errors but we used a binomial model for analysis.