Supplemental Material

All model specifications

Main model specification

All reported results come from the use of a generalized linear mixed effects model that included transition score, triplet membership, position, and exposure duration as fixed effects. Exposure duration was coded as an unordered factor as we had no *a priori* predictions that representations would necessarily differ in a linear way across short, medium, and long exposure. The model also included fixed pairwise interaction terms for transition score and triplet membership, and triplet membership and position. Interactions of each transition score, triplet membership, and position with exposure duration were also included as fixed effects. By-subject random slopes and intercepts were also included for transition score, triplet membership, and position, and the pairwise interactions between transition score and triplet membership, and triplet membership and position. Because exposure duration was a between-subjects factor, it was not included in any random effects terms. Thus, the final model was coded, using the lme4 package in R (Bates, Machler, Bolker, & Walker, 2015) as follows:

```
glmer(Old ~ Transition Score + Triplet Membership + Position + Exposure Duration +
        (Transition Score : Triplet Membership) + (Triplet Membership : Position) +
        (Transition Score : Exposure Duration) + (Triplet Membership : Exposure Duration) +
        (Position : Exposure Duration) +
        (Position + Triplet Membership + Transition Score + (Position : Triplet Membership) +
        (Triplet Membership : Transition Score) || Subject))
```

For our control group, who received random exposure to six presentations of each shape, exposure duration was not included in the model because it was not manipulated for these participants. The model was otherwise identical.

Model specification for shuffled vs. triplet sequences model

In our preregistered analyses, we found that participants rely on only triplet membership to guide memory decisions after short exposure (and not transition score or position). To follow-up on this and further interrogate the idea that early representations rely more on a general sense of what goes together, independent of order, than later representations, we compared old responses to ordered vs. unordered triplets from exposure. Specifically, we ran an additional model using only the triplet sequences (which were generally excluded from the main paper) and test sequences which had a triplet membership score of 3 but did not maintain the original order (e.g., BCA, henceforth "shuffled sequences"). As discussed in the main paper (*Early emergence of group knowledge*), if there is an initial step in which order is irrelevant, these two test types should be treated as equally old initially and become increasingly distinct with extended exposure. The model was specified as follows:

```
glmer(Old ~ Test Type + Exposure Duration +
        (Test Type : Exposure Duration) +
        (Test Type || Subject))
```

Model specification for assessing effects of test third in structured exposure groups

In addition to the control study described in the main text, we also wanted to confirm that the memory judgements of participants who were exposed to structured input (across all exposure durations) did not change in a systematic way over the course of test. To do this, we ran a model that included third of the test phase as a fixed effect. For these analyses (described in detail below, see *Confirming main effects are present throughout test*), third of test was included in the model as a numeric variable (1, 2, or 3). Because third was a within-subjects manipulation it was also included in the random effects. The pairwise interactions were not included as random effects due to failure of this more complicated model to converge. Thus, the final model was as follows:

```
glmer(Old ~ Transition Score + Triplet Membership + Position + Third +
        (Transition Score : Triplet Membership) + (Triplet Membership : Position) +
        (Transition Score : Third) + (Triplet Membership : Third) + (Position : Third) +
        (Position + Triplet Membership + Transition Score + Third || Subject))
```

Control Analyses

Results do not change when exposure triplets are included in analysis

While our test sequences were orthogonalized along each dimension as much as possible, the exposure triplets represent the highest possible value on each dimension. Thus, out of an abundance of caution, and to be sure that our results were not driven by the extreme values associated with exposure triplets, in the main text we reported only models that interrogate variability in behaviour across test sequences that were not exposure triplets. We believe the results presented in the main paper better reflect the independent contributions of each factor. This model (now run on all test sequences) was otherwise identical to the main model described above. Like in the main paper, we report main effects using the Anova function from the Car package (Fox & Weisberg, 2019).

Results from this model largely mirror the results from the models run excluding exposure triplets; there were significant main effects of transition score and triplet membership, such that increasing the transition score (χ^2_1=23.67, p < .001) or triplet membership score (χ^2_1=15.22, p < .001) significantly increased the likelihood that participants would endorse a test sequence as old. As when excluding the triplets, there was also no main effect of position when they were included (χ^2_1 = 1.63, p = 0.20), indicating that increasing the number of items that maintained their position relative to exposure did not increase the likelihood that someone would endorse a test sequence as old.

There was also a main effect of exposure duration (χ^2_1 = 7.64, p = 0.02): Pairwise comparisons indicated that the participants were more likely to make an old response after the short exposure duration as compared with after medium (β = 0.42, *SE* = 0.15, z = 2.76, p = .006, Odds Ratio = 1:1.53). There was no difference between medium and long exposure (β = 0.19, *SE* = 0.15, z = 1.23, p = 0.22, Odds Ratio = 1:1.21), nor between long and short (relevelled model with long exposure as reference level; β = 0.23, *SE* = 0.15, z = 1.53, p = 0.12, Odds Ratio = 1:1.26). Qualitatively, short exposure participants were more likely to respond “old”, medium exposure participants were not very likely to respond “old”, and long exposure participants endorsed sequences as “old” to an intermediate degree. This is somewhat in contrast to our findings presented in the main text (when exposure triplets were not included), where the difference in old responses between short and long exposure was also significant. However, because the analysis presented here collapses across old responses to both triplets and foils, we do not think examining the overall level of old responses will provide additional, meaningful information, beyond the results of our three factors of interest.

When all sequences were included there was no longer a significant interaction between triplet membership and transition score (χ^2_1 = .0006, p = 0.995). There was, however, a significant interaction between triplet membership and position (χ^2_1 = 11.83, $p < .001$), such that position mattered more when there were more items from one triplet present in the same test sequence. The fact that this interaction was not present when triplet sequences (see main paper, *Both specific transitions and general groupings contribute to memory judgements*) were excluded suggests it was driven by "old" responses to triplet sequences, and reflects a difference in how participants respond to exposure triplets and foil sequences which include three items from one triplet in a shuffled order (ABC vs. CAB, for example). (Note that this finding and interpretation are reminiscent of the result that shuffled foils and triplets are more confusable after short exposure than after extended exposure, which was presented in the main text and is detailed below; see SI, *Shuffled and triplet sequences less distinct after short exposure*.)

Finally, all interactions between exposure duration and each factor of interest remain essentially unchanged relative to the main text. There was a significant interaction between transition score and exposure duration (χ^2_2 = 10.99, p = .004). Pairwise comparisons again indicated that transition score mattered less after short exposure than after medium exposure (β = -0.33, SE = 0.10, z = -3.24, p = 0.001, Odds Ratio = 1:0.72) or long exposure (relevelled model with long exposure as reference level: β = -0.23, SE = 0.10, z = -2.21, p = 0.03, Odds Ratio = 1:0.80), but that there was no difference in the extent to which transition score impacted old-new judgements between medium and long exposure (β = -0.11, SE = 0.10, z = -1.03, p = 0.30, Odds

Ratio = 1:0.90). Neither the interaction between triplet membership and exposure duration ($\chi^2_2 = 0.31$, $p = 0.86$) nor position and exposure duration ($\chi^2_2 = 0.42$, $p = 0.81$) was significant. This suggests that our finding that item-item transition knowledge emerges to shape old-new memory judgements between six and 30 repetitions of each triplet is still true when exposure triplets were also included in the model.

Assessing differences across exposure duration groups in overall performance

Our main questions of interest in this paper were 1) whether position, transition score, and/or triplet membership were represented in memory following statistical learning, and 2) whether these factors came to guide memory at different points during the learning process. While we thus included exposure duration to examine the second of these questions, it is also possible that manipulating exposure duration resulted in overall performances differences between our groups. In order to ensure that our results are not attributable to differences in overall performance (additional fatigue with additional exposure time, for instance), we compared response times and exclusion rates between each of our exposure duration groups.

Response times. We first compared the test-phase reaction times across exposure conditions, reasoning that if participants in the long exposure condition were particularly fatigued they would respond more slowly than participants in other exposure duration conditions. Because participants in all exposure duration conditions completed the same number of test trials, previous exposure duration is the only source of difference across groups. The results of a mixed effects model which predicted reaction time from exposure duration (with by-subject random effects) showed no significant difference in the reaction times between any pair of exposure durations (all p's > 0.48).

Exclusion rates. Increased fatigue in the longer exposure duration condition could also manifest as a larger number of participants failing our attention checks. Importantly, all of our original analyses excluded participants who failed more than one third of attention checks, reducing the concern that participants who contributed to our analyses were not paying attention. That said, if fatigue were plaguing our long exposure participants more than our medium or short exposure duration participants, we might expect more long exposure subjects to be excluded due to inattention. The exclusion rates did not follow this pattern: Instead, the medium exposure condition had the greatest number of participants excluded (N=48) followed by short (N=30) and then long (N=21). Only the difference between medium and long was statistically significant (48 vs. 21, $\chi^2_1 = 7.16$, $p = 0.01$); other pairwise comparisons were not significant (30 vs. 48, $\chi^2_1 = 2.56$, $p = 0.11$; 30 vs. 21, $\chi^2_1 = 0.94$, $p = 0.33$).

Taken together, these results fail to find evidence of significantly more fatigue in the long exposure group, suggesting that general differences between the groups' responses are not likely to be adversely impacting our results.

Confirming main effects are present throughout test phase

In the main paper, we presented results confirming that our effects of interest did not change over the course of the test after structured exposure (see *No evidence of learning during test after structured exposure*) to demonstrate that the observed memory behaviour arose from learning phase experience. Beyond confirming that our effects of interest did not *change* with test phase experience, we also wanted to confirm that the results were qualitatively similar *within* the thirds of test, and that each exposure duration group was impacted similarly by test experience.

Across exposure durations, main effects qualitatively present from first third of test. Because our test phase included one presentation of each test sequence per third of test, the data from the first third of our test represent the cleanest possible measure of exposure-phase only learning in that they avoid the possible confound of repeated presentations at test. Thus, we ran an additional model which included only the first presentation of each test sequence (collapsed across all exposure durations to account for much more limited data). Similar to our main results, restricting analyses to the first third also showed that transition score and triplet membership were marginally predictive of old responses (transition score, $\chi^2_1 = 3.29$, p = 0.07; triplet membership, $\chi^2_1 = 3.80$, p = 0.05), while position was not ($\chi^2_1 = 2.40$, p = 0.12), suggesting that our effects were present during the first third of the test phase (albeit a bit weaker) and did not require additional test phase repetitions to emerge.

Minimal effect of test third in each exposure duration. Separate from whether our main effects were present from the earliest test, it is also possible that test third might have differentially impacted each exposure duration group. Thus, as an additional exploration of how test phase may contribute to our results, we also investigated the effect of test third in each exposure duration separately. There was an interaction between triplet membership and third in the medium exposure group ($\beta = 0.12$, $SE = 0.03$, $z = 2.22$, $p = 0.03$, Odds Ratio = 1:1.13), such that triplet membership more strongly predicted old-new judgements towards the end of test, which we will discuss further below. We saw no other significant interactions between third and any factor of interest in the short, medium, or long exposure groups (all β's < abs(0.16), all p's > 0.11).

To further investigate how third of test was influencing triplet membership judgements in the medium exposure condition, we ran smaller models on each third's data separately. These models included only the main effects of triplet membership, position, and transition score, because model comparison for this limited data set indicated that the more complex model did not provide a significant improvement above the simpler model (BIC Main Effects Model = 2204.3, BIC Interaction Model = 2215.6, $X^2(2) = 3.58$, $p = 0.17$). The results of these models indicated that transition score and triplet membership were both significant predictors of old-new judgments in all test thirds for the medium exposure group (all transition score β's > 0.30, p's < 0.02; all triplet membership β's > 0.17, p's < 0.01). This suggests that although the strength of triplet membership representations increased over third of test for our medium exposure group (as noted in the paragraph above), triplet membership and transition score were significant predictors of old-new judgements from the first third onward. Position did not impact old-new judgements in any third (all β's < abs(0.15), p's > 0.15).

In summary, we saw no other significant interactions between test third and any factor of interest in the short, medium, or long exposure groups as mentioned above (all β's < abs(0.16), all p's > 0.11). Thus, the only metric which changed over the course of test in any exposure group was triplet membership, which more strongly predicted old-new judgements as the test went on in the medium exposure group. Even in this group, triplet membership predicted old-new judgements from the first third of test, and just became a stronger predictor as participants gained additional test-phase experience. We take this to mean that, overall, there is very little evidence that our measures change across the time spent taking the test, as would be expected if participants' representation of structure were predominantly formed during this phase of the experiment. Particularly when considered together with the null effects in our control group (who watched a random, unstructured presentation of images prior to test), our results suggest that participants' old-new judgements are very likely reflecting representations formed during the exposure phase.

Bidirectional item-item links do not predict responses better than forward transitions

We chose to operationalize our measure of item-item transitions as the sum of the forward TPs for each pair of items in order to both align our work with past research and maximally differentiate our three metrics. Thus, while the bulk of our analyses define item-item links as the strength of the *forward* transition between two items, it is also possible that *bi-directional* links between two items—for the sake of this example, A and B—are being stored instead. If this is the case, a metric of bi-directional transitions that considers "AB" and "BA" to be equally old should better fit our data than one considering forward transitions alone.

To investigate this possibility, we calculated a "neighbouring score" for each of our test items, which reflected the relative frequency with which the constituent bigrams (pairs) co-occurred, irrespective of order, during the exposure phase. For example, test sequence ABC (an exposure triplet) and test sequence CBA (a shuffled sequence) both received a neighbouring score of 2 (AB/BA pair, and BC/CB pair), a test sequence such as ACB received a score of 1 (CB pair), and a test sequence such as AIE received a score of 0.5 (since IA could have appeared in half the transitions between triplets at exposure). We then ran our original model using this neighbouring score in the place of our transition score to examine whether neighbouring better predicted our pattern of results.

First, it is important to note that a BIC-based model comparison suggested that our original transition score model fit our data significantly better than this neighbouring score (BIC Transition Score Model = 17,762, BIC Neighbouring Model = 17,790, $\chi^2(20) = 28.09$, $p <$ 0.001). This suggests that our original forward transition score measure (in combination with our triplet membership score) better accounted for our data than did a neighbouring score. Nevertheless, the results using neighbouring were virtually identical to those using our original transition score. As in the main paper, triplet membership and neighbouring score both positively predicted the likelihood of endorsing an item as old (Triplet membership: $\chi^2{}_1 = 33.95$, $p < 0.001$; Neighbouring score: $\chi^2{}_1 = 5.17$, $p = 0.02$), while position had no impact on old-new judgements ($\chi^2{}_1 = 1.34$, $p = 0.25$). There was also a main effect of exposure duration ($\chi^2{}_1 = 5.75$, $p = 0.056$), although this was marginally weaker than in the main results. Like in the transition score analyses, triplet membership and neighbouring score also interacted ($\chi^2{}_1 = 11.75$, $p = 0.001$), as did exposure duration and neighbouring score ($\chi^2{}_1 = 6.05$, $p = 0.048$). Finally, there was an interaction between position and triplet membership score which was not present in the main paper ($\chi^2{}_1 = 4.83$, $p = 0.03$).

One reason that our transition score-based model may be a better fit than this neighbouring score is specific to our task design: For our stimulus set, neighbouring score is more correlated with triplet membership score than is transition score. Because unordered grouping is reflected in our triplet membership score (which awards test sequences like CBA with a high score, similar to the neighbouring metric), test sequences which were low in their transition score but high in this neighbouring score were also sequences which best orthogonalized our transition score and triplet membership score. This also helps to explain why the interaction between neighbouring score and exposure duration was smaller than between transition score and exposure duration: After short exposure when triplet membership is largely driving old-new decisions, the neighbouring score also reflects triplet membership information (and thus the difference between short and medium exposure duration is less drastic).

Do memory judgements track with binary "seen" status?

One potential concern with our transition metric is that, by necessity in our design, it is correlated with what participants had seen during exposure. Given this, it is not fully possible to say whether or not the memory behaviour explained by our transition score is reflecting an accumulated record of all the transitions which had been seen previously, or whether participants had calculated a transitional probability between each pair of items which they could tap into at test. For example, some of our test sequences (BCD, for example) reflect a series of events which participants had in fact seen, despite not being exposure triplets. It is possible that participants' "old" responses to this type of test sequence could be driven by a tendency to respond "old" to all sequences they have seen, rather than having learned the strength of individual transitions. While we confirmed that forward transitions better predict memory than bidirectional associations (and thereby provided some support for the idea that a record of all item pairs does not explain our data as well as our transition score, see SI, *Bidirectional item-item links do not predict responses better than forward transitions*) we nevertheless were interested in further clarifying that our effects could not be more parsimoniously explained by participants responding "old" to test sequences they had seen previously.

Memory behaviour is not predicted by seen vs. unseen metric. First, we were interested in understanding whether our effect of transition score was attributable to participants responding "old" to entire test sequences they had seen previously. To examine this, we translated our transition score into a binary score indicating whether or not the particular test sequence had been seen previously. We reasoned that if the effects of transition score were attributable to test sequences with higher transition scores having been seen during exposure, the seen vs. unseen metric should predict responses at least equally as well as the original transition score. Under this coding scheme "part-word" foils (like "DAB") received a score of 1; all other foil sequences were scored as 0 (and, as in our original analyses, triplets were not included). We then ran our original model but replaced the transition score metric with this new seen vs. unseen variable. However, this new model indicated that our seen vs. unseen metric was not a significant predictor of old responses ($\chi^2_1 = 0.79$, $p = 0.37$). Importantly, the other results from this model were very similar to those reported in the main paper, including our triplet membership score which was still a significant predictor of old responses ($\chi^2_1 = 28.55$, $p < 0.001$).

Triplet membership results hold when omitting all "seen" sequences. Independent of the whether our effects of transition score could be attributed to what participants had seen previously, we wanted to confirm that our other metrics of interest (triplet membership, in particular) were unaffected by this potential confound. As such, a second important control analysis was to demonstrate our other results of interest also hold when excluding sequences that have been seen. As such, we also ran a model on the subset of our data that excluded any "seen" sequence. Crucially, the results of this model indicate that group membership was a significant predictor of old-new judgements across exposure durations ($\chi^2_1 = 13.19$, $p = 0.00002$), despite greatly reduced data. This highlights that group membership predicts old-new decisions even when omitting test sequences that have been seen before.