Supplemental Information

Analysis S1. Individual differences

Experiment 1. We examined whether individual differences in working memory capacity, learning rate across array repetitions, and long-term memory performance were associated with one another. We operationalized working memory capacity using an independent change detection task. We operationalized learning rate using the individual subject slopes from the linear mixed model with random slopes and random intercepts. Finally, we quantified long-term memory performance as d' in the final recognition test. The independent measure of working memory capacity correlated robustly with both long-term memory performance, $r = 0.51$, $p < .001$ and with learning rate in the repeated whole-report task, $r = .48$, $p < .001$. The relationship between learning rate and long-term memory performance was significant but numerically less robust, $r = .30$, $p = .03$. Together, the individual differences analyses suggest that individual differences in initial working memory capacity, learning rate, and long-term memory performance were all interrelated in this experiment. However, some of these relationships may require larger sample sizes to replicate reliably as between-subject correlations require higher sample sizes than our main within-subjects effects of interest (Baker et al. 2020). In particular, the reliability of the long-term memory measure (d': 0.19, hit rate: 0.43, false alarm rate: 0.37) was lower than either of the two working memory measures (change detection K: 0.75, mean number correct: 0.97), in part because of lower trials counts for the long-term memory recognition task.

Experiment 2. In Experiment 2, we found that most, but not all, pairs of measures correlated with one another. Specifically, although the learning rate in the repeated whole-report task correlated with both long-term memory performance, $r = .30$, $p = .03$, and with the independent measure of working memory capacity, $r = .34$, $p = .01$, we did not find a significant correlation between long-term memory performance and the independent measure of capacity, $r = .09$, $p = .51$. Thus, although all three measures were indirectly linked via the learning rate in the repeated whole-report task, we did not see significant interrelationships between all possible pairs of measures. This may be caused by lower statistical reliability of the long-term memory measure in this experiment (e.g., overall, d-prime values were closer to floor in this task) or because of random noise due to the sample size. As in Experiment 1, the reliability of the long-term memory measure (d': 0.33, hit rate: 0.35, false alarm rate: 0.46) was lower than either of the two working memory measures (change detection K: 0.81, mean number correct: 0.97). Our primary effects of interest were within-subjects effects that we are well-powered to detect (Baker et al. 2020; Xu et al. 2017), but between-subjects correlations are presented here for completeness.

Analysis S2. Characterizing whether participants recalled items in a consistent order.

To quantify how consistently participants reported items across all repetitions of an array, we computed transition probabilities between consecutive responses throughout the learning period. To make the logic of the analysis clear, we will give a concrete example. If participants reported the 6 array items in the same order across all 8 repetitions of the learned array, we would get a 1x42 vector like this: [1 2 3 4 5 6 1 2 3 4 5 6 1 2 3 4 5 6 …]. So, if we see that the participant clicked item 1, then there would be a 100% transition probability to 2, and a 0% transition probability to any other number. If participants reported in a random order each time, the vector that we created may look something like this: [1 2 3 4 5 6 2 3 1 5 3 6 4 1 6 5 3 2 …]. Now, the transition probability starting from 1 is no longer predictable.

We measured the order in which participants responded to the items within an array, and we calculated the transition probabilities of the responses across learning of the array (a vector of 42 responses, i.e., 6 item responses x 8 repetitions). From this vector, we computed a transition probability matrix. In Experiment 1, there are a total of 36 transition possibilities (6 x 6 items, e.g., 1->1, 1->2, 1->3, … 2->1, 2->2, 2->3, etc). To make the responses to arrays comparable to one another, we vectorized and rank-ordered the probabilities from each transition probability matrix (1 matrix was computed for each of the 30 unique arrays). For example, in Array #1, there may have been a 90% probability of 1->2 and a 0% probability of 1->4. However, in Array #2 there may have instead been a 90% probability of 1->4 and a 0% probability of 1->2. By rank ordering the probabilities for each array from high to low, we can then average across them. For each participant, we ended up with an N array x 36 transition probabilities matrix. After examining transition probabilities across the entire learning period (Figure S1A), we also separately computed transition probabilities for the first half of learning (Repetitions 1-4) and the last half of learning (Repetitions 5-8) for each array (Figure S1B). Finally, we computed two key baseline conditions of interest. First, in the random response condition, we simulated a random response order for each repetition of the array. We imposed similar constraints as were required by the task (i.e., participants must respond to all 6 items only 1 time each during each repetition of the array). Second, in the 100% consistent condition, we simulated a perfectly consistent response order for all 8 repetitions of the array.

We found that participants make responses in an order that is much more consistent than would be expected by chance (gray line). The highest two transition probabilities were notably high (>90%). A transition probability of 100% would indicate that the participants made adjacent responses in a consistent order across all 8 repetitions. Therefore, the empirical pattern that we observed would be consistent with an account in which participants form links between the first three items starting on Repetition 1 (i.e., two transition probabilities, Item 1->2 and Item 2->3), and then accumulate a consistent response order for the remaining items at later repetitions. Consistent with this notion, we separately analyzed transition probabilities for the first half of repetitions (Repetition 1-4) and the last half of repetitions (Repetition 5-8), and we found that participants' responses were significantly more consistent for the later repetitions of an array. We repeated these analyses for Experiment 2 and found similar results (Figure S2).

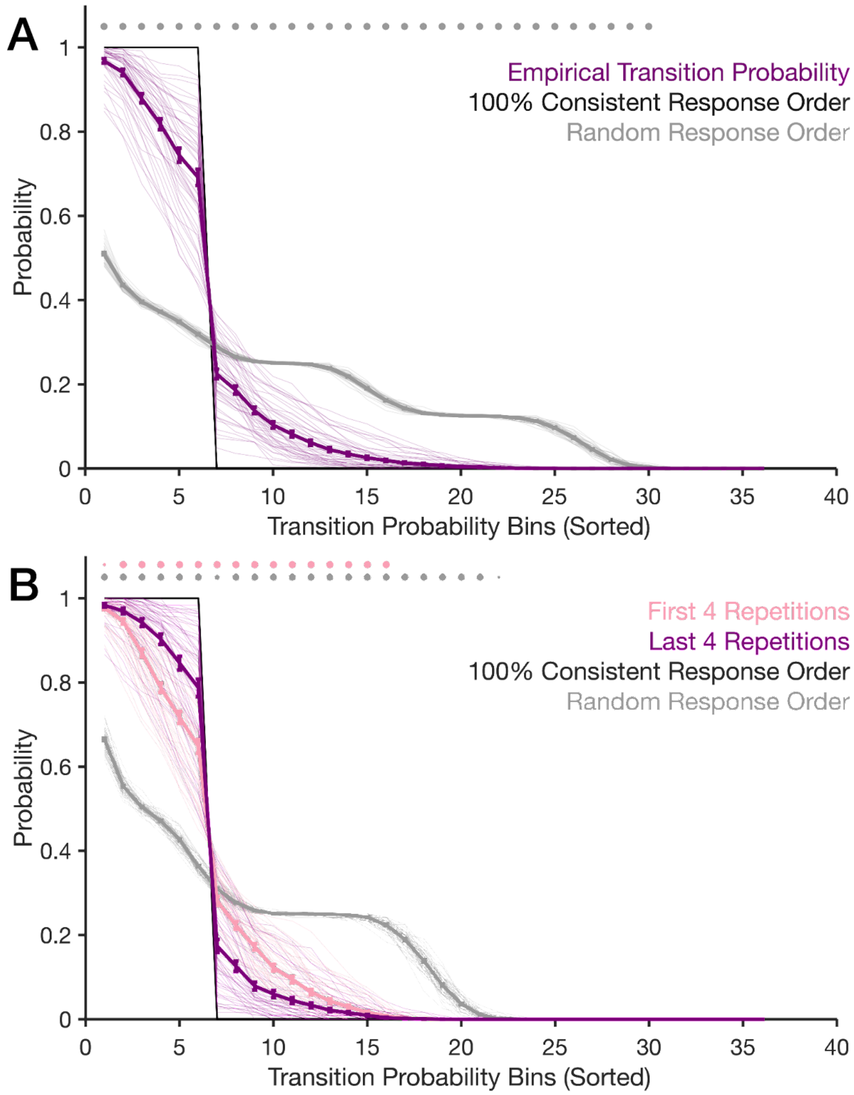

Figure S1. Transition probability between responses in Experiment 1. (a) Transition probability in the order of reporting items across the 8 repetitions of an array. The purple line represents the empirical data. The black line indicates a baseline for a 100% consistent response order across all 8 repetitions. The gray line represents responding to items in a random order on each repetition of the array. Gray dots indicate a significant difference between the purple and gray lines. (b) Transition probabilities for the first half of learning (Repetitions 1-4) versus the last half of learning (Repetitions 5-8) for each array. The pink line represents the empirical data for Repetitions 1-4. The purple line represents the empirical data for Repetitions 5-8. The black line indicates a baseline for 100% consistent responding across all eight repetitions. The gray line represents responding in a random order on each repetition of the array. Gray dots indicate a significant difference between the purple and gray lines. Pink dots indicate a significant difference between the purple and pink lines. In both panels, error bars represent +/- 1 standard error of the mean. The size of the dots represents significance thresholds (large p < .001, medium p<.01, small p < .05).

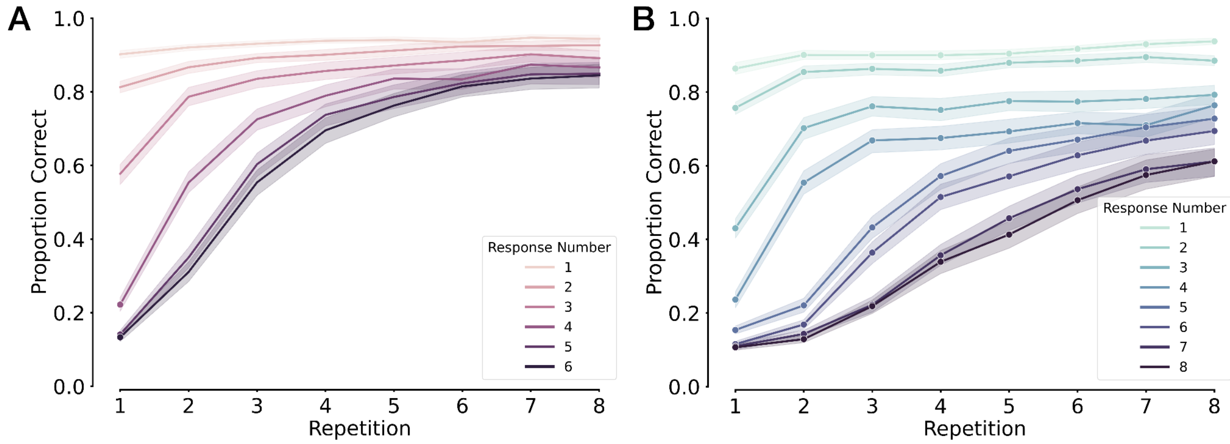

Figure S2. Accuracy as a function of response number and repetition. (A) Accuracy in Experiment 1 as a function of response number and repetition. (B) Accuracy in Experiment 2 as a function of response number and repetition.

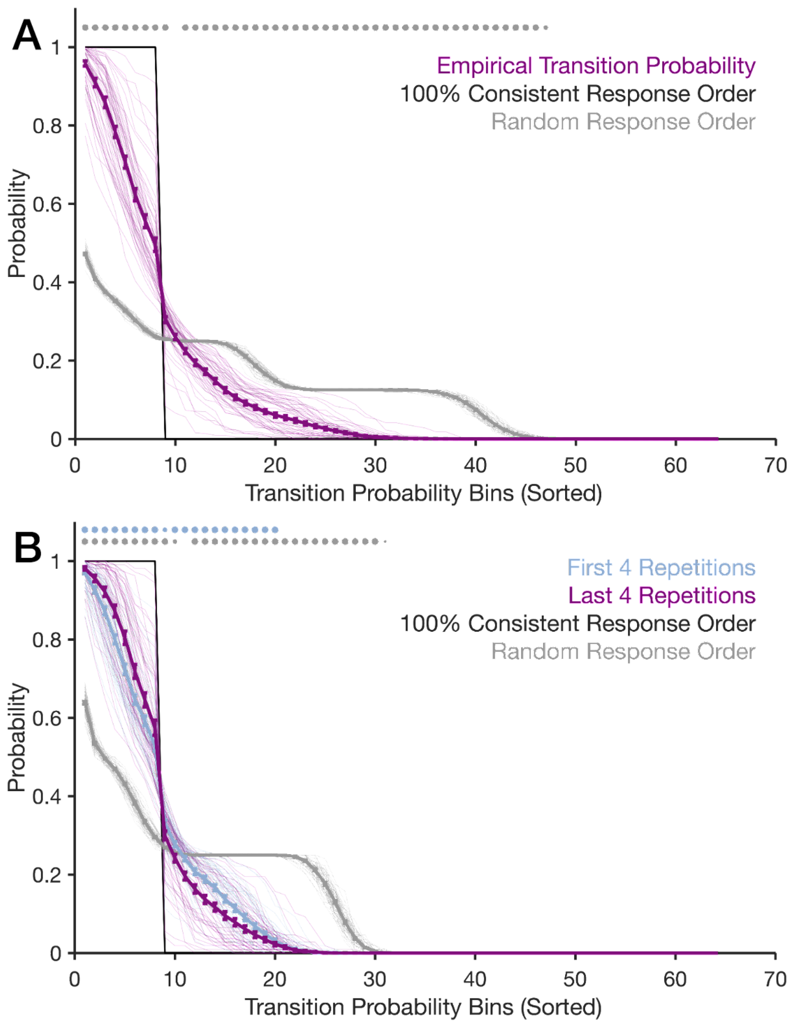

Figure S3. Transition probability between responses in Experiment 1. (a) Transition probability in the order of reporting items across the 8 repetitions of an array. The purple line represents the empirical data. The black line indicates a baseline for 100% consistent responding. The gray line represents responding in a random order on each repetition of the array. Gray dots indicate a significant difference between the purple and gray lines. (b) Transition probabilities for the first half of learning (Repetitions 1-4) versus the last half of learning (Repetitions 5-8) for each array. The blue line represents the empirical data for Repetitions 1-4. The purple line represents the empirical data for Repetitions 5-8. The black line indicates a baseline for 100% consistent responding across all eight repetitions. The gray line represents responding in a random order on each repetition of the array. Gray dots indicate a significant difference between the purple and gray lines. Blue dots indicate a significant difference between the purple and blue lines. In both panels, error bars represent +/- 1 standard error of the mean. The size of the dots represents significance thresholds (large p < .001, medium p<.01, small p < .05).

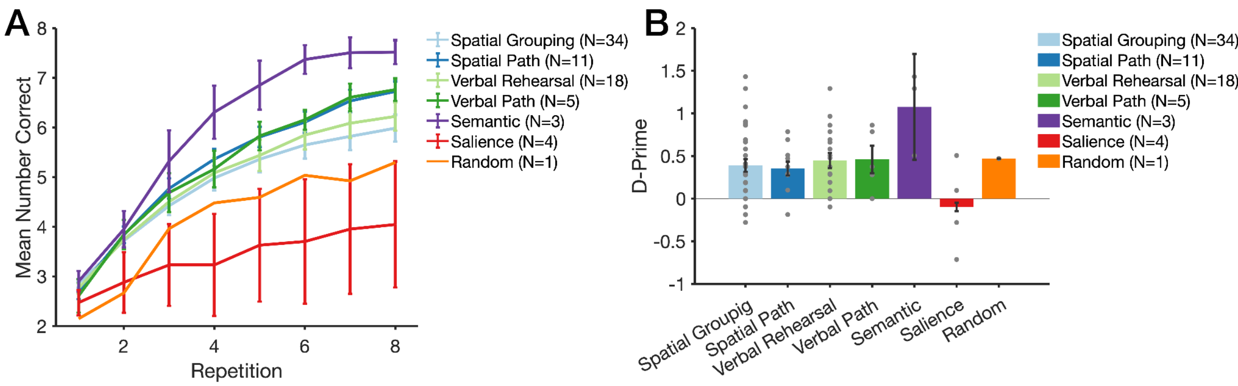

Figure S4. Performance as a function of strategy. Participants answered an open-ended question asking to describe the strategy they used to perform the repeated whole-report task. Participants' responses were scored by 3 raters as being consistent or not consistent with 7 strategy categories. (A) Performance in the main repeated whole-report task as a function of coded strategy. (B) Performance on the long-term memory recognition task as a function of coded strategy. Relative to the modal strategy (Spatial Grouping, N=34), these results are tentatively consistent with semantic coding being a more effective strategy than spatial grouping ($p = .01$) and salience (i.e., encoding brightest/darkest first) being a poorer strategy than spatial grouping ($p = .04$). However, future work specifically targeting strategy use is needed given the limitations of this sample, including (1) very small sample sizes when the results are broken out by strategy and (2) many participants reporting the use of more than one strategy.

References

Baker, Daniel H., Greta Vilidaite, Freya A. Lygo, Anika K. Smith, Tessa R. Flack, André D. Gouws, and Timothy J. Andrews. 2020. "Power Contours: Optimising Sample Size and Precision in Experimental Psychology and Human Neuroscience." *Psychological Methods*, July. https://doi.org/10.1037/met0000337.

Xu, Zhan, Kirsten C. S. Adam, Xinyi Fang, and Edward K. Vogel. 2017. "The Reliability and Stability of Visual Working Memory Capacity." *Behavior Research Methods*. https://doi.org/10.3758/s13428-017-0886-6.