Supplementary Materials for "Working Memory Shapes Information Sampling and Attention Allocation Across Development"

Anonymous

Blind for Review

Supplementary Materials for "Working Memory Shapes Information Sampling and Attention Allocation Across Development"

Chance Level Calculation for Priority Score in Experiment 2

To determine the appropriate chance level, we calculated the probability that feature D would appear in the first or second sampling position by chance. Given that participants could sample 1 to 7 features on each trial, we make a simplifying assumption of equal probability (i.e., $\frac{1}{7}$) of sampling each sample size (i.e., 1 to 7 features). We therefore computed: (a) the probability that k features are sampled ($\frac{1}{7}$ for each value of k), (b) the probability that D feature is in the set of k (i.e., $\frac{k}{7}$), and (c) $\frac{2}{k}$ (the probability the D feature was sampled first or second).

- When $k = 1$, the probability that is the D feature is: (a) $= \frac{1}{7} \times$ (b) $= \frac{1}{7} \times$ (c) $= 1 = \frac{1}{49} = 0.02$.
- When $k = 2$, the probability the D feature is sampled is: (a) $= \frac{1}{7} \times$ (b) $= \frac{2}{7} \times$ (c) $= 1 = \frac{2}{49} = 0.04$
- When $k = 3$, the probability the D feature is sampled, and it was sampled first or second is: (a) $= \frac{1}{7} \times$ (b) $= \frac{3}{7} \times$ (c) $= \frac{2}{3} = 0.04$.
- When $k = 4$, the probability the D feature is sampled, and it was sampled first or second is: (a) $= \frac{1}{7} \times$ (b) $= \frac{4}{7} \times$ (c) $= \frac{2}{4} = 0.04$.
- When $k = 5$, the probability the D feature is sampled, and it was sampled first or second is: (a) $= \frac{1}{7} \times$ (b) $= \frac{5}{7} \times$ (c) $= \frac{2}{5} = 0.04$.
- When $k = 6$, the probability the D feature is sampled, and it was sampled first or second is: (a) $= \frac{1}{7} \times$ (b) $= \frac{6}{7} \times$ (c) $= \frac{2}{6} = 0.04$.
- When $k = 7$, the probability the D feature is sampled, and it was sampled first or second is: (a) $= \frac{1}{7} \times$ (b) $= 1 \times$ (c) $= \frac{2}{7} = 0.04$.

Summing it all up, results in $p \approx 0.265$.

Justification for using 0.5: We used 0.5 as a conservative threshold that substantially exceeds the calculated chance level of 0.265. We use this conservative approach to mitigate the risk that our simplifying assumption that all values of k (i.e., the number of sampled features) are equiprobable.

Learning and Sampling Pattern in Experiment 2

Further examination of learning trajectories reveals two types of learners: Rapid Learners who achieved perfect accuracy (11 consecutive correct trials, $p < .05$ relative to chance) in less than 11 trials and General Learners who achieved the same performance later. There were 34 Control Rapid Learners and 1 Control Gradual Learner and 15 Dual-Task Rapid Learners and 13 Dual-Task General Learners. The remaining 7 Dual-Task adults did not meet this criterion. Figure S1A shows these different Learner Types' accuracy during the training phase aligned by their learning point, the first trial of the 10 consecutive correct trials. The Control group was also collapsed into one Learning type for simplification.

If the dual-task manipulation only shifted participants to rely on another learning system that was minimally impaired, we would observe the same number of features sampled between the Dual-Task and Control conditions. Particularly, Rapid Learners who relied on hypothesis-testing-based learning should only sample the rule feature. However, as shown in Figure S1B, both Rapid Learners and General Learners from the Dual-Task conditions sampled significantly more features than the Control participants in the post-learning period (median difference$_{Rapid-Control} = 0.82$, $p = .03$, $d = 0.68$, median difference$_{general-Control} = 2.01$, $p < .001$, $d = 1.29$, from permutation tests). While we cannot rule out the possibility of slower learning driving broad sampling, the persistent broader sampling in Dual-Task Rapid Learners supports the direct impact of WM on attention and information sampling independent of learning trajectory.

Learning and Sampling Pattern in Experiment 3

Similar to Experiment 2, we examined whether the distributed attention observed in the Dual-Task condition resulted from reliance on a qualitatively different, non-selective learning system. We identified two learning types based on performance trajectories: Rapid Learners, who achieved high accuracy (at least 17 correct out of 20 consecutive trials, $p < .05$ relative to chance) within the first 10 trials, and General Learners, who reached this criterion later in training. Among Control adults, 34 of 36 were classified as Rapid Learners, with the remaining 2 failing to meet either criterion. The Dual-Task group had 20 Rapid Learners, 11 General Learners, and 3 participants who did not meet either criterion (Figure S2A).

Crucially, post-learning attention patterns revealed persistent differences between conditions regardless of learning trajectory. As shown in Figure S2B, Dual-Task participants maintained broader attention distribution than Control participants even after successful learning, with General Learners showing significant differences (median difference = 0.13, $p = .026$, $d = 1.00$) and Rapid Learners showing marginally significant differences (median difference = 0.11, $p = .024$, $d = 0.78$) in permutation tests. This finding is particularly important: even Dual-Task participants who exhibited rapid, hypothesis-testing-like learning continued to sample more broadly than Controls. This persistent difference in attention distribution, independent of learning trajectory or success, provides strong evidence that WM constraints directly affect information sampling rather than simply shifting participants to an alternative learning system.

GCM Model Specification in Experiment 3

As described in the main text, we simulated the recognition memory from participants's categorization performance with the generalized context model (GCM) by Nosofsky (1986).

The GCM outlines that for two exclusive categories A and B, the probability of classifying Stimulus S_i in Category A (C_A), denoted as $P(R_A|S_i)$, is given by:

$$P\left(R_A \mid S_i\right) = \frac{b_A \sum_{j \in C_A} \text{sim}_{ij}}{b_A \sum_{j \in C_A} \text{sim}_{ij} + (1 - b_A) \sum_{k \in C_B} \text{sim}_{ik}} \tag{1}$$

where b_A ($0 \leq b_A \leq 1$) is the response bias for Category A, and sim_{ij} and sim_{ik} represent the similarities between a given exemplar i and exemplars in categories A and B, respectively. The similarity is an exponential decay function of psychological distance d, shown in:

$$\text{sim}_{ij} = \exp\left[-c d_{ij}\right] \tag{2}$$

where c ($0 \leq c \leq \infty$) is a scaling parameter indicating discriminability in the psychological space. This parameter can be influenced by task and individual variables, such as stimulus noise, stimulus exposure time, multitasking, or individual perceptual or memory acuity Nosofsky and Zaki (2002). Psychological distance d is calculated using the standard Euclidean formula:

$$d_{ij} = \left[\sum w_m \left(x_{im} - x_{jm}\right)^2\right]^{1/2} \tag{3}$$

where x_{im} is the psychological value of exemplar i on dimension m, and w_m ($0 \leq w_m \leq 1$, $\sum w_m = 1$) is the attentional weight allocated to dimension m. These weights are free parameters, capturing the amount of attention allocated to each dimension during categorization(Nosofsky, 1986). Increased attention to a dimension enhances its discriminability (and potentially memorability), while reduced attention has the opposite effect. The GCM posits that recognition judgments are based on the cumulative similarity of an item to all exemplars across categories, denoted as fam, as seen in:

$$\text{fam}_i = \sum_{i=1}^{k} sim_{ik} \tag{4}$$

The recognition response rule is given by:

$$P\left(Old \mid S_i\right) = \frac{fam_i}{fam_i + \beta} \tag{5}$$

where a lower β value indicates a stronger tendency to respond "Old".

We estimated attentional weights of D and P features from responses on Switch items ($P_{flurp}D_{jalet}$ and $P_{jalet}D_{flurp}$) using Equations 1–3, assuming equal response biases for both categories and setting c as 2.50. To simulate recognition performance, we used Equations 2–5, estimating the probability of correct "New" responses to the new-D item and the one-new-P item for every possible β value between 20 and 0.1 in 0.1 increments. Lower β values indicate a greater tendency to recognize items as "old", even when familiarity is low.

References

Nosofsky, R. M. (1986). Attention, similarity, and the identification-categorization relationship. *Journal of Experimental Psychology. General*, *115*(1), 39–61. doi: 10.1037//0096-3445.115.1.39

Nosofsky, R. M., & Zaki, S. R. (2002). Exemplar and prototype models revisited: Response strategies, selective attention, and stimulus generalization. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *28*(5), 924–940. Retrieved 2022-12-01, from `http://doi.apa.org/getdoi.cfm?doi=10.1037/0278-7393.28.5.924` doi: 10.1037/0278-7393.28.5.924

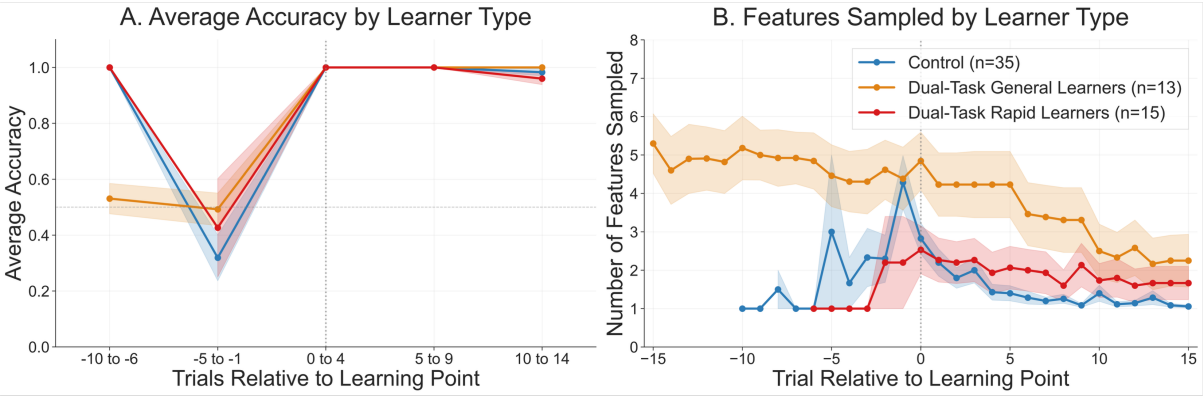

Figure S1. A. Average Accuracy by Learner Type. B. Number of Features Sampled by Learner Type. Learning point is the first trial of 10 consecutive correct trials. Error band stands for the standard error of the means.

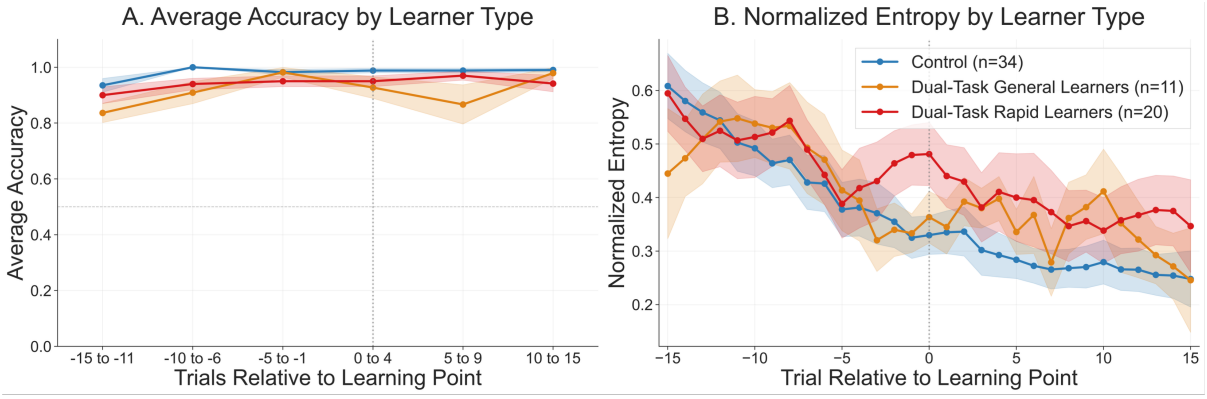

Figure S2. A. Average Accuracy by Learner Type. B. Normalized Entropy by Learner Type. Learning point indicates the first trial of the 17-out-of-20 correct sequence. Error bands represent standard errors of the means.