Supplemental Materials for

Children's Long-Term Retention is directly constrained by their Working Memory Capacity Limitations

1. Sample Size Determination

We used Bayes Factors design analysis (BFDA; Schönbrodt & Stefan, 2018) with 10000 simulations to test our ability to detect group differences or their absence for our most central measure, the LTM/WM ratio, using a non-directional between-subjects Bayesian t-test (using a Bayes Factor > 3 as a decision criteria), with a minimum number of 30 participants per age group, and an additional 10 x 2 participants per group if evidence was inconclusive. We later realized that we had incorrectly entered the total samples sizes (rather than the per group estimates required for the BFDA *t.between* test), for the between-group comparison reported in the pre-registration. The proportion of studies correctly showing evidence for (or against) group differences would be lower than the simulations reported in the pre-registrations, but we did reach our boundary of BF > 3 at *N* = 40 per group and stopped data collection as planned.

We also simulated our ability to find evidence against a correlation, first assuming that there is no correlation in the population (*r* = 0), at the lowest possible sample size (total *N* = 4 groups × 30 participants = 120). 91.7% of samples showed evidence for H_0, 7.6% were inconclusive, 0.7% showed evidence for H_1. Finally, assuming an effect of *r* = 43 (based on the smallest correlation in Exp. 1b, in Vogel & Fukuda, 2019), 98.6% of samples showed evidence for H1, 1.3% were inconclusive, and 0.1% showed evidence for H0.

2. Detailed Pre-Registered Exclusion criteria

We will exclude data from participants who:

1) Have a total of ten or more missing values from the WM and LTM task (indicating a within-study ‘breaks’ longer than 10 minutes, because it might indicate that they were distracted whilst performing the task, or failure to complete the task/technical problems).

2) Perform close to floor level (< .55) for the smallest set size (two items) in the WM task, which we interpret as an indication that participants were not paying attention to the task (or did not understand the task).

3) Participants who responded ‘studied’ to > 90% of novel items in the LTM task, which might indicate that they did not understand the task.

One participant in the youngest age group was excluded and replaced for performing close to floor level at set size 2.

3. Detailed inclusion criteria and demographic information

Participants were included based on the following criteria:

- native speaker of English
- nationality must be British, American or Canadian
- normal or corrected-to-normal vision
- no cognitive impairment or dementia
- no language-related disorders
- age must be between 6 and 13 years, or 18 – 30.

Also, the online consent form will state the following **ELIGIBILITY REQUIREMENTS:**

• Must be 6 – 13 or 18 – 30 years of age with normal or corrected-to-normal vision and hearing including normal color vision meaning the ability to distinguish shades and tint of color

• Must be fluent in the English language

- Must not have a diagnosis of photosensitive epilepsy in which seizures are triggered by flashing lights; bold, regular patterns; or regular moving patterns

Demographic information is presented in Table S1 below.

4. Results when excluding all participants with negative *p*(WM) or *p*(LTM) values

In the paper, we presented results where negative *p*(WM) or *p*(LTM) values were adjusted to theoretically plausible values, such that *p*(WM) values equivalent to holding < 1 item in WM were adjusted to equal 1, and *p*(LTM) values < 0 were adjusted to 0. Below, we report the output of the second approach to such values, in which participants were completely excluded if they had a < 0 value at one or more set sizes, either for *p*(WM) or *p*(LTM). Using this latter approach, 4 observations of < 0 *p*(WM) values and 30 *p*(LTM) values resulted in the exclusion a total of 25 participants (9 of the youngest children, 6 of the second youngest, 4 adolescents, and 6 adults), some of whom had more than one < 0 value, leaving n=135 in the present analysis.

We explored how many of the items encoded into WM could be retrieved on an LTM test (i.e., the LTM/WM ratio). First, we compared the average ratios of the very youngest child group with the adults and found weak evidence against age differences in the average LTM/WM ratio (BF_{01}=2.36; youngest children average ratio: *M*=0.38, *SD*=0.20, *N* = 31, young adults, *M*=0.44, *SD*=0.14, *N* = 34). Next, we found strong evidence that the LTM/WM ratio differed by Set Size ($BF_{10}=1.89\times10^{12}$, *F*(1, 397)=49.63, η_p^2=0.111), evidence against an age group difference (BF_{01}=8.78, *F*(3, 397)=1.83, η_p^2=0.014) and against an age group × set size interaction (BF_{01}=28.06, *F*(3, 397)=1.22, η_p^2=0.009). At set size 2 (1st-2nd Graders: *M*=0.43, *SD*=0.17; 3rd-4th Graders: *M*=0.53, *SD*=0.22; 5th- 7th Graders: *M*=0.57, *SD*=0.12; Adults: *M*=0.55, *SD*=0.16) set size 4 (1st-2nd Graders: *M*=0.38, *SD*=0.27; 3rd- 4th Graders: *M*=0.38, *SD*=0.19; 5th- 7th Graders: *M*=0.41, *SD*=0.21; Adults: *M*=0.40, *SD*=0.20) and set size 6 (1st-2nd Graders: *M*=0.33, *SD*=0.33;

3rd- 4th Graders: *M*=0.31, *SD*=0.19; 5th- 7th Graders: *M*=0.35, *SD*=0.25; Adults: *M*=0.36, *SD*=0.23). Note that, when using the adjusted values, we found some evidence *for* an interaction. We discuss this difference in the main manuscript.

We found ‘decisive’ evidence that LTM memory performance (*p*(LTM) was better for items presented for lower-set size items ($BF_{10=}3.36 \times 10^{54}$, $F(1, 397)=262.78$, $\eta_p^2=0.398$) and evidence for an age group effect ($BF_{10=}226.23$, $F(3, 397)=13.67$, $\eta_p^2=0.094$), and evidence against an interaction ($BF_{01=}73.32$, $F(3, 397)=1.49$, $\eta_p^2=0.011$).

Finally, we observed a positive correlation between average age-standardized *k*-scores and *p*(LTM) scores, $r=.31$, $BF_{10}=1.75\times10^2$. Next, we tested separate, age-standardized *k*-scores and *p*(LTM) correlations for memory from items presented as part of Set Sizes 2, 4 and 6, respectively (Set Size 2: $BF_{10}=7.60$, $r=.22$, Set Size 4: $BF_{10}=2.24$, $r=.18$, and Set Size 6: $BF_{10}=10.92$, $r=.24$).

5. Working Memory Accuracy

First, we tested whether our WM set size manipulation had the intended effect on WM performance. We used hierarchical Bayesian logistic regression (implemented in the *brms* package for *R*, Bürkner; 2017, 2018; R Core Team, 2020). For this analysis, we considered responses marked as guesses as incorrect. We compared a model with main effects of age groups (as a categorical factor) and set size (a continuous factor), to a model including these main effects, and a set size and age group interaction. The BF in favor of the model without the interaction was 1.05×10^3 over a model not including this factor. We report the output of the favored model.

We found credible evidence that memory performance decreased as set size increased (η=-0.51; SE=0.02, 95% Bayesian Credible Interval; BCI[-0.55,-0.48]). There was also credible evidence for differences between the youngest children and the second youngest children (η=0.42; SE=0.16, 95% BCI[0.10,0.73]), the adolescents (η=0.67; SE=0.16, 95% BCI[0.37,0.98]), and the adults, η=1.05; SE=0.16, 95% BCI[0.73,1.36]).

6. Confidence ratings

See Figure S1 for the confidence Ratings (1-6, indicating both the *same* versus *different* judgment and the confidence).

Table S1. *Participant Demographics by Age Group (in percentages)*

	1st – 2nd graders	3rd – 4th graders	5th– 7th graders	Adults
Race				
American Indian / Alaska Native	-	-	-	2.5
Asian	20.0	17.5	2.5	7.5
Black or African American	-	12.5	5.0	10.0
More Than One Race	-	-	-	2.5
Native Hawaiian or Other Pacific Islander	-	-	-	-
White (or European)	80.0	67.5	8.5	75.0
Other	-	-	-	2.5
Prefer not to say	-	2.5	5.0	-
Unknown	-	-	2.5	-
Ethnic				
Hispanic or Latino/Latina	2.5	2.5	7.5	5.0
Not Hispanic or Latino/Latina	80.0	72.5	75.0	80.0
Other	17.5	25.0	12.5	15.0
Prefer not to say	-	-	5.0	-

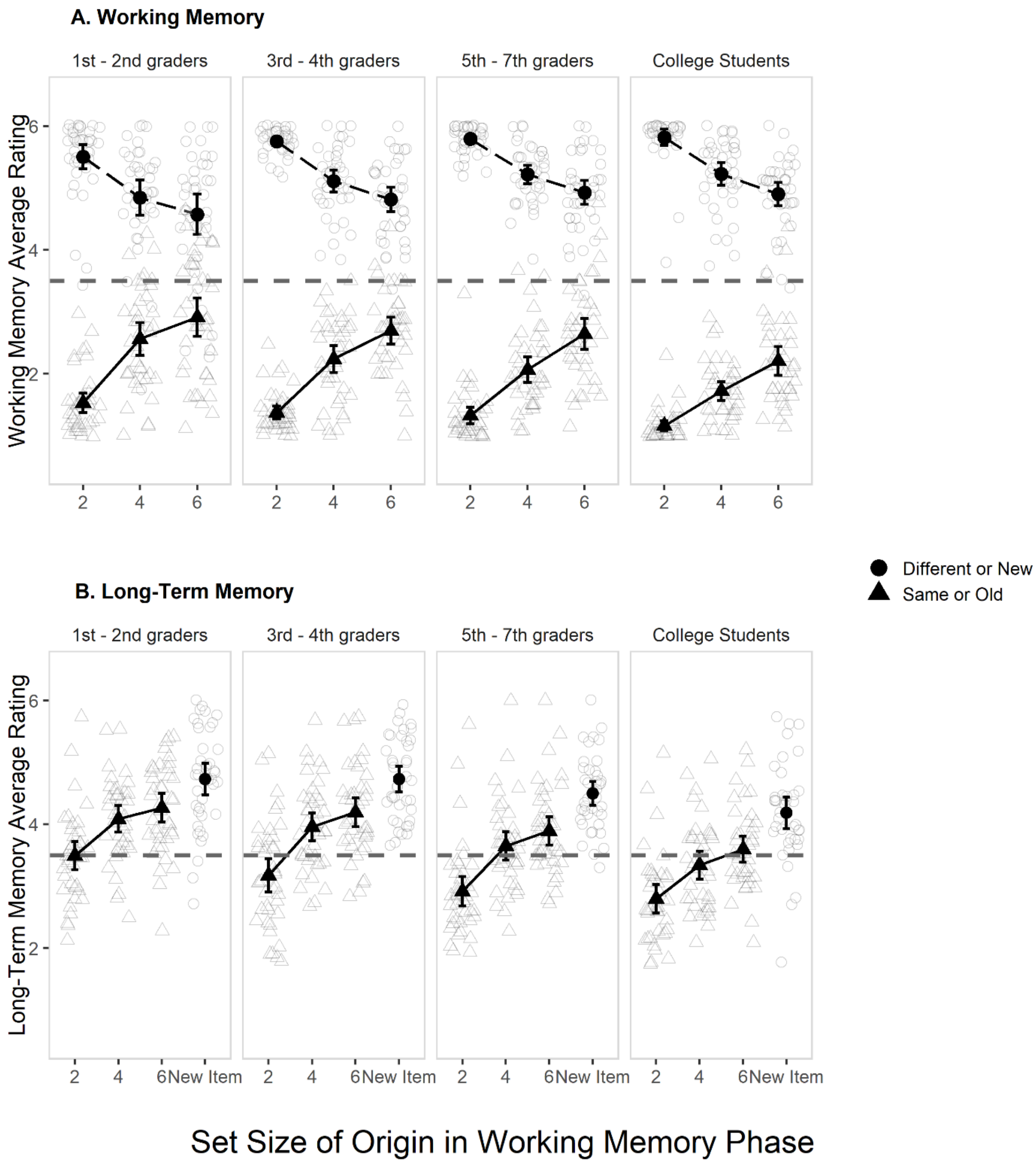

Figure S1. Average ratings (1–6; 1=*sure same/studied*, 2=*believe same/studied*, 3=*guess same/studied*, 4=*guess different/new*, 5=*believe different/new*, 6=*sure different/new*) by set size in the working memory (WM) task. **(A)** WM task ratings; **(B)** long-term memory (LTM) task ratings. Circles show ratings on trials when the probe item was different or new, and diamonds show performance when the probe item was the same as one in the studied set, or old. The black circles and diamonds represent the overall mean ratings. The transparent, smaller circles and diamonds represent individual participants' mean ratings. These are jittered to avoid overlap. Error bars on the group means represent 95% confidence intervals. The dashed, horizontal lines represent the neutral point (Rating = 3.5), such that any data points above that line represent predominantly “new” responses and any points under it, predominantly “old” res