

Supplementary Online Materials for:

The Temporal Dynamics of the Pretesting and Prequestioning Effects

This document contains: (a) analyses of retention interval timing, (b) analyses of errors made on the retention test, and (c) log-transformed power function analyses.

Analyses of Retention Interval Timing

Participants in all delayed retention interval conditions had to complete the retention test within 24 hours of the scheduled release time, and completion within 12 hours was incentivized via a bonus payment. To explore potential variability in actual completion times across retention interval conditions, we calculated completion times relative to the first observed completion time within each retention interval group. Whenever a retention test was released online, which was always timed to coincide exactly with the intended retention interval, at least some participants began completing it almost immediately. Hence, these completion times are a reasonable proxy for participants' actual completion times. As shown in Table S1, mean completion times were often between 1 and 5 hours and did not show clear systematic variation across groups, except the 3-day condition of Experiment 2 which had a mean completion time of 6.37 hours.

Table S1

Retention Test Completion Time Relative to the Earliest Completion Time per Retention Interval

Experiment	Retention interval	Mean completion time after earliest recorded completion time, in hours	<i>SD</i> , in hours
1	1-day	2.87	5.00
	3-days	2.46	4.10
	7-days	1.23	2.74
	14-days	4.12	5.17
2	1-day	3.18	4.16
	3-days	6.37	7.90
	7-days	2.77	4.73
	14-days	2.98	3.62
	30-days	4.67	6.29

To further address the potential impact of completion times for the delayed retention interval groups, we performed supplementary analyses for both experiments using the same procedures as detailed in the main text—a mixed-design ANOVA and a three-parameter power function fit—on data excluding the subset of participants with retention test completion times of more than 10 hours past the first observed completion time for a given retention interval.

The supplementary analyses for Experiment 1, which excluded 10 participants, yielded the same overall pattern of results. The ANOVA revealed a significant interaction between learning method and retention interval, $F(4, 181) = 7.69$, $MSE = 0.02$, $p < .001$, $\eta_p^2 = 0.15$. Post-hoc comparisons indicated significantly higher recall in the pretesting than the reading condition at 5 min, 1 day, and 3 days, but not at 7 or 14 days. For the power function analyses, likelihood-ratio tests comparing the parameters a (initial performance) and b (forgetting rate) between the pretesting and reading conditions showed that parameter a differed significantly between conditions, $\chi^2(1) = 94.93$, $p < .001$ ($\Delta a = 0.23$, 95% CI [0.18, 0.27]), whereas parameter b did not, $\chi^2(1) = 1.62$, $p = .203$ ($\Delta b = 0.07$, 95% CI [-0.03, 0.18]).

The supplementary analyses for Experiment 2, which excluded 26 participants, also yielded similar overall patterns of results. The ANOVA revealed a significant interaction between learning method and retention interval, $F(9.83, 414.69) = 2.82$, $MSE = 0.04$, $p = .002$, $\eta_p^2 = 0.06$. Post-hoc comparisons indicated significantly higher recall in the prequestioning than the studying condition at 5 min, 3 days, and 7 days, but not at 1 day ($p = .0516$), 14 days, or 30 days, as well as significantly higher recall in the prequestioning than the no-activity condition at 5 min, 1 day, 3 days, and 7 days, but not at 14 or 30 days. For the power function analyses, likelihood-ratio tests comparing the parameters a (initial performance) and b (forgetting rate) showed that parameter a differed significantly between the prequestioning and studying conditions, $\chi^2(1) = 11.88$, $p < .001$ ($\Delta a = 0.14$, 95% CI [0.06, 0.22]), as well as between the

prequestioning and no-activity conditions, $\chi^2(1) = 42.86, p < .001$ ($\Delta a = 0.28$, 95% CI [0.22, 0.36]), whereas parameter b did not differ significantly between any pair of conditions ($\chi^2(1) \leq 1.40, p \geq .236$).

Analyses of Retention Test Errors

To explore the nature of the retention test errors made across conditions and retention intervals, one author coded each incorrect response for potential classification into each of two categories: *related* responses, which were incorrect answers that nonetheless closely resembled the correct answer in spelling or phonological form, and *factually plausible* responses, which were incorrect answers that did not closely resemble the correct answer but could plausibly complete the underlying fact or statement. The coding rubric that was used is archived on OSF. To determine the reliability of the coding procedure, a second independent rater scored the same data using the same procedures, yielding good reliability (Cohen's $\kappa \geq 0.87$).

Descriptive statistics by error type for Experiments 1 and 2 are shown in Tables S2 and S3, respectively, as follows. Overall patterns were broadly stable across retention intervals and conditions in both experiments, with no consistent systematic variation in either error type.

Table S2

Retention Test Errors in Experiment 1

Error type	Retention interval	Proportion of incorrect responses	
		Pretesting	Reading
Related	5-min	0.091	0.093
	1-day	0.022	0.072
	3-days	0.055	0.060
	7-days	0.066	0.075
	14-days	0.063	0.060
Factually plausible	5-min	0.539	0.573
	1-day	0.716	0.715
	3-days	0.726	0.772
	7-days	0.757	0.787
	14-days	0.786	0.789

Table S3*Retention Test Errors in Experiment 2*

Error type	Retention interval	Proportion of incorrect responses		
		Prequestioning	Reading	No activity
Related	5-min	0.118	0.128	0.084
	1-day	0.030	0.072	0.041
	3-days	0.023	0.066	0.022
	7-days	0.050	0.036	0.072
	14-days	0.041	0.081	0.027
	30-days	0.051	0.009	0.046
Factually plausible	5-min	0.503	0.618	0.559
	1-day	0.685	0.713	0.688
	3-days	0.642	0.728	0.723
	7-days	0.705	0.770	0.760
	14-days	0.813	0.759	0.765
	30-days	0.749	0.753	0.776

Log-Transformed Power Function Analyses

To examine the robustness of the curve-fitting results to data transformation, the power function analyses were repeated using a log-based (geometric) transformation of the observed recall rates. As detailed below, the transformed analyses yielded results that closely matched those obtained using the original (untransformed) recall data, with no meaningful differences in model fit or parameter estimates. These findings suggest that the observed patterns are not sensitive to the choice of transformation.

In Experiment 1, the model using a log-based transformation of the observed recall rates with a common scaling parameter ($c = 0.05$) provided an adequate fit for the two conditions (pretesting, reading), $\chi^2(3) \leq 5.00$, $p \geq .172$, $R^2 \geq .972$, with results closely matching those from the original fitting approach (parameter a : $\chi^2(1) = 9.94$, $p = .002$, $\Delta a = 0.30$, 95% CI [0.10, 0.60]; parameter b : $\chi^2(1) = 0.070$, $p = .791$, $\Delta b = 0.11$, 95% CI [-5.53, 6.21]).

In Experiment 2, the corresponding model ($c = 0.96$) also provided an adequate fit for the

three conditions (prequestioning, studying, no activity), $\chi^2(4) \leq 6.00$, $p \geq .199$, $R^2 \geq .970$, again yielding results consistent with the original fitting approach (prequestioning vs studying: $\chi^2(1) = 3.60$, $p = .058$, $\Delta a = 0.61$, 95% CI [-0.24, 1.67]; prequestioning vs no activity: $\chi^2(1) = 13.41$, $p < .001$, $\Delta a = 0.91$, 95% CI [0.27, 1.91]; parameter b : $\chi^2(1) \leq 1.80$, $p \geq .180$, $\Delta b \leq 0.27$). It should be noted that 95% CIs are based on bootstrap resampling with log-transformed re-aggregation; estimates for b are less stable under this specification due to nonlinear transformation and joint re-estimation.