Supplementary materials

Paper: The benefits of memory control processes in working memory: comparing effects of self-reported and instructed strategy use

Bartsch, Souza & Oberauer

Content

1. **Serial position curves**

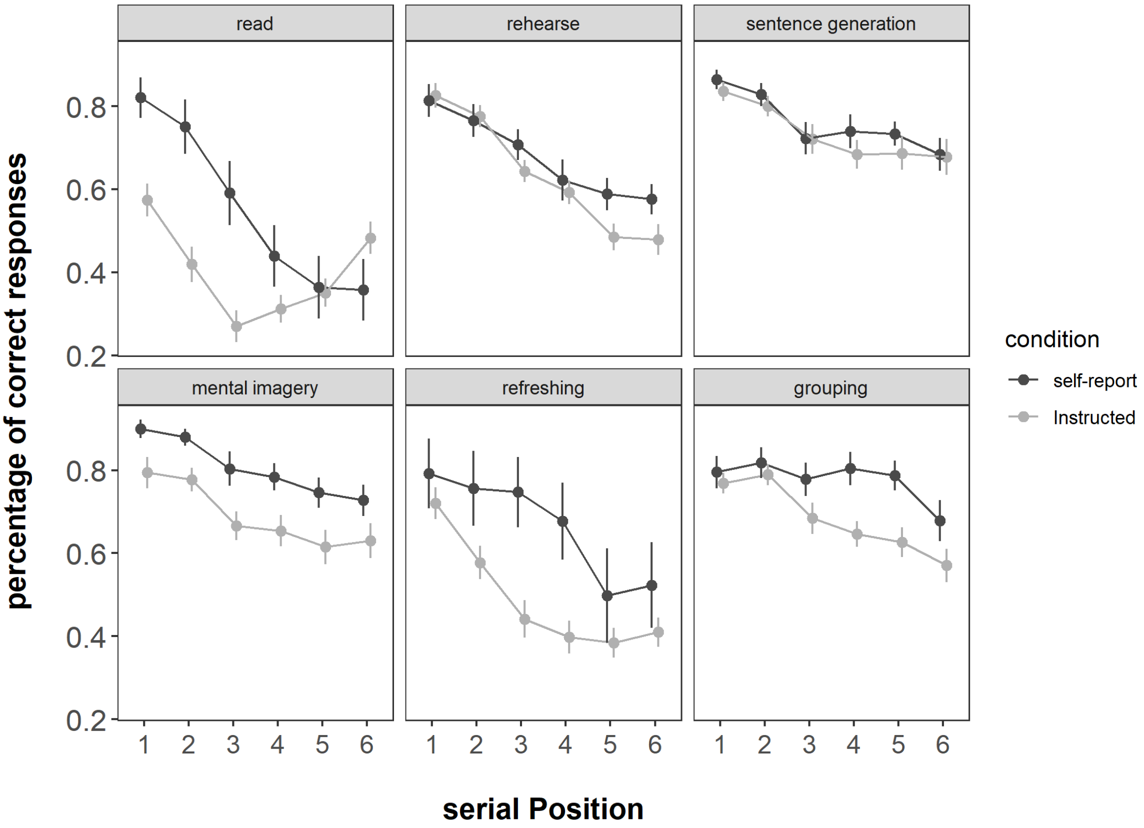
We had no a-priori hypothesis about the interaction of serial position with the effect of instructed compared to self-chosen strategies on immediate recall performance. For the interested reader we present the data and analysis of serial position curves here. It is to note that the serial-position curves of the grouping strategies do not show the common saw-tooth pattern found with temporal grouping (i.e., mini serial-position curves within the groups). This is likely because we are averaging over subject's idiosyncratic grouping patterns (some group in 3s, others in 2s, other in mixed group sizes), so that pattern washes out.

We further present the immediate serial recall performance broken down by concreteness below (see **Figure S3**). For Experiment 1, we included word concreteness as a factor in the analysis, but that factor had only a main effect, without entering into any interactions, and therefore we decided to drop it for Experiment 2 and present participants with mixed lists of abstract and concrete words for simplicity.

1.1. Experiment 1

Figure S1

Immediate serial recall performance across serial position, strategies and session (self-report vs. instructed) in Experiment 1. Error bars represent 95% between-subjects confidence intervals.



We analyzed the immediate recall data using Bayesian generalized linear mixed models (BGLMM) implemented in the R package *brms* (Bürkner, 2017, 2018). BRMS – as opposed to the BayesFactor package - allows the inclusion of continuous factors as fixed effects. As serial position is such a continuous factor, we decided to implement the analysis differently to the ones presented in the main manuscript. With this, we model serial position as a linear effect, meaning we do not model the U-shape of the curve, but the primacy effect. We chose to do so, as we were not interested in testing the existence or shape of the serial position curve (which is beyond dispute) but the question whether the condition contrast is larger at the beginning or the end of the list. This translates into an interaction of condition with the linear contrast of serial position.

The dependent variable was the binary accuracy (i.e., correct or incorrect) of recalled responses. Correct responses were defined as recalling the target item at the correct serial position. Therefore, we assumed a Bernoulli data distribution predicted by a linear model through a logit link function (i.e., a repeated-measures logistic regression). The fixed-effects were serial position (1 - 6), session (instructed vs. self-reported), and strategy (read vs. refreshing vs. rehearsal vs. grouping vs. mental imagery vs. sentence generation), as well as their interactions. Following the recommendation of Barr and colleagues (Barr, Levy, Scheepers, & Tily, 2013; see also Schielzeth & Forstmeier, 2009) we implemented the maximal random-effects structure justified by the design; by-participant random-intercept and by-participant random-slope for serial position, strategy, and condition.

The regression coefficients were given Cauchy priors with a scale of 0.353. These scales were chosen because these priors were recently proposed as default priors for model comparisons with logistic models (Oberauer, 2019).

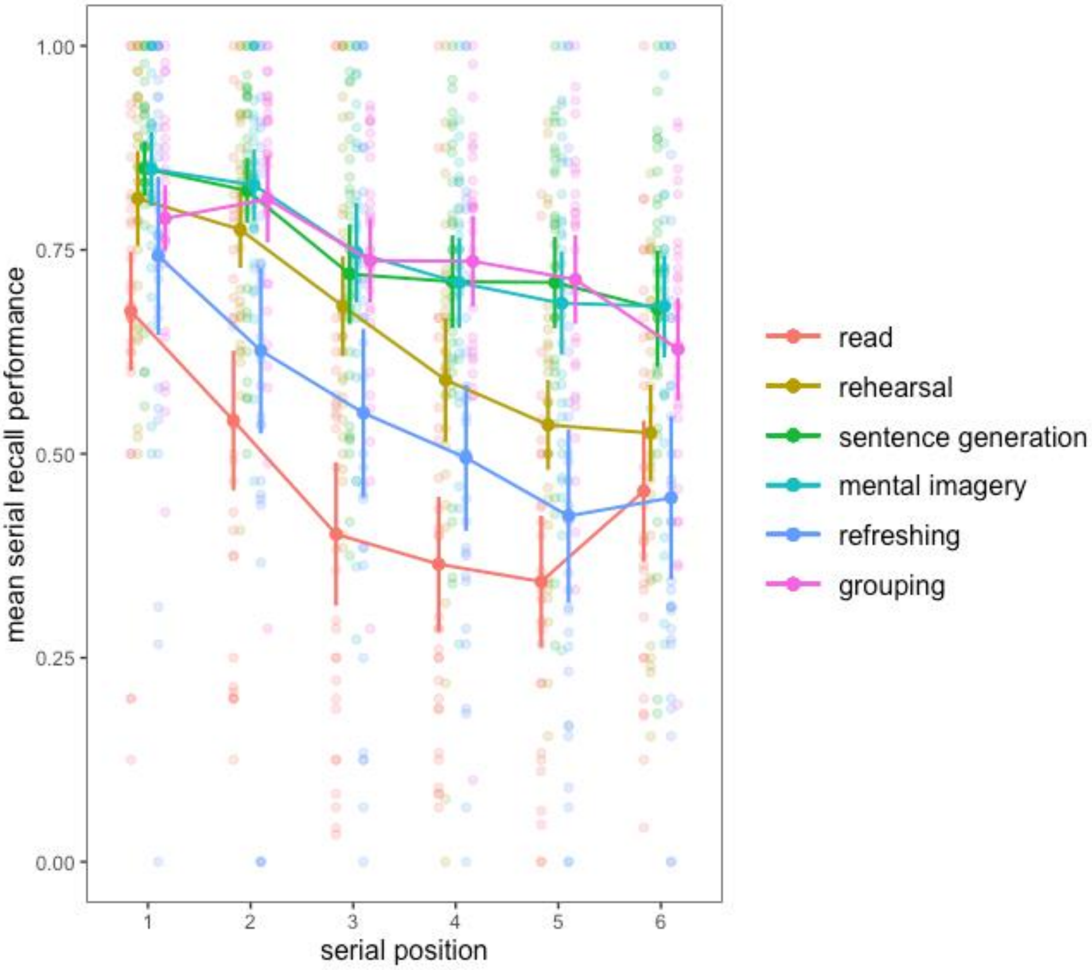
We used an MCMC algorithm (implemented in Stan; Carpenter et al., 2017) that estimates the posteriors by sampling parameter values proportional to the product of prior and likelihood. These samples are generated through 4 independent Markov chains, with 2000 warmup samples each, followed by 50000 samples drawn from the posterior distribution which are retained for analysis. Following Gelman and colleagues (2013), we confirmed that the 4 chains converged to the same posterior distribution by verifying that the R-hat statistic – reflecting the ratio of between-chain variance to within-chain variance – was < 1.01 for all parameters, and we visually inspected the chains for convergence. Finally, we used the *bayes_factor* function in the *brms* package, which implements the bridge sampler (Gronau, Singmann, & Wagenmakers, 2020), for computing the BFs.

1.1.1 Results Experiment 1

The generalized mixed effects model on the data of Experiment 1 revealed strong evidence *against* the three-way interaction effect of serial position x session x strategy (BF_{01} = 1235). There was evidence for the two-way interaction of serial position by strategy (BF_{10}= 1.77× 10^{6}), with the differences in immediate serial recall performance across the different strategies being credible for all serial positions except for the first (SP1 BF_{01}= 0.77; see **Figure S2**). There was evidence as well for the interaction of session by strategy [BF_{10}= 3.52]). There was no evidence for the serial position by session interaction [BF_{10}= 0.80]. Finally, all main effects were credibly supported by the data (strategy [BF_{10}= 1.24 x 10^{108}]; session [BF_{10}= 1.03x10^{17}]; serial position [BF_{10}= 6.44x10^{46}]).

Figure S2

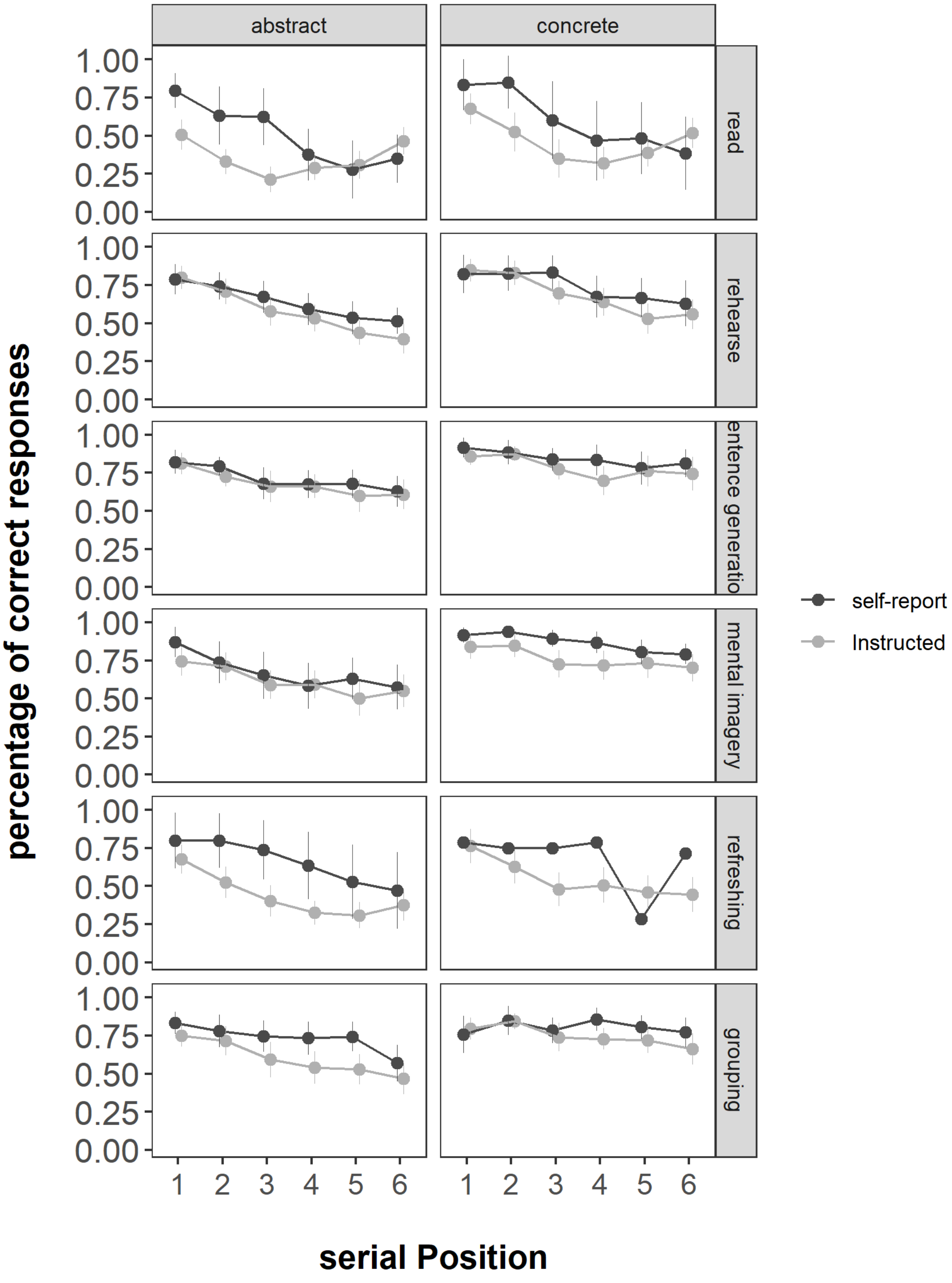
Immediate Serial Recall performance in Experiment 1 for each serial position and strategy (averaged over sessions). Small dots represent the data of individual subjects. Error bars show the 95% within-subject confidence interval.



1.1.2 Serial position curves by concreteness

Figure S3

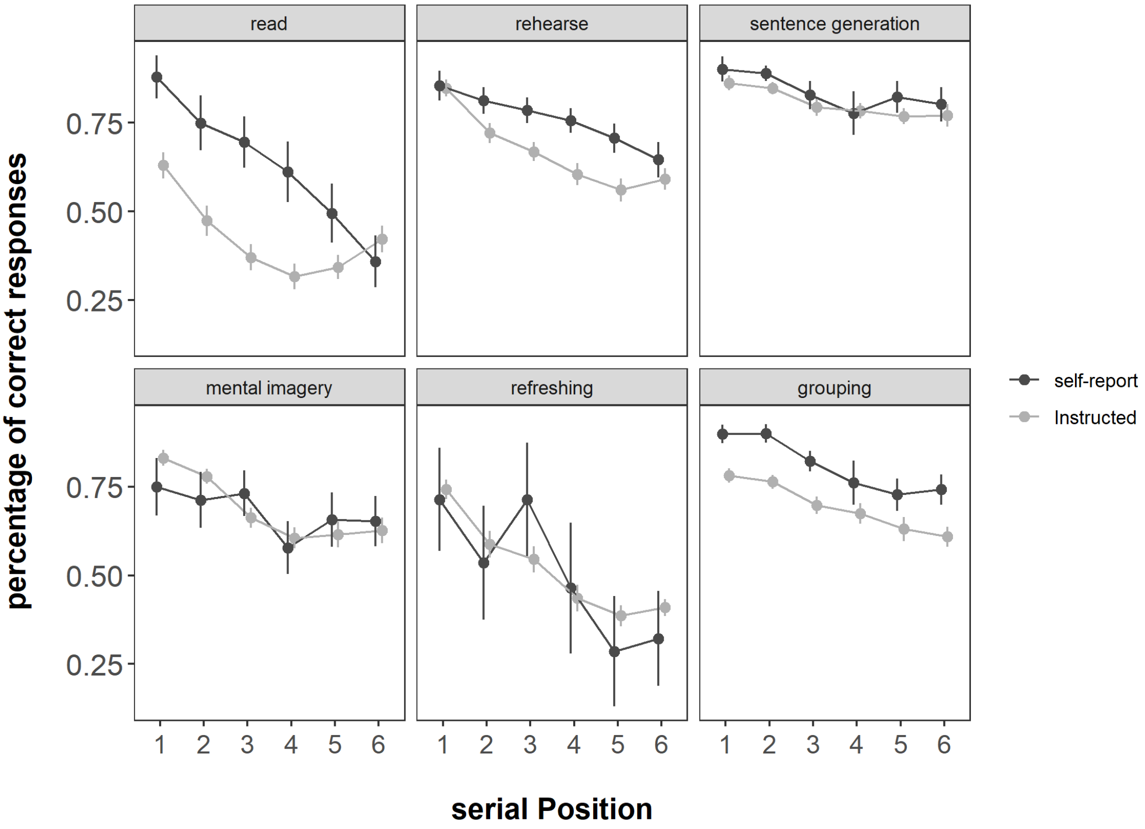
Immediate serial recall performance across serial position, strategies, concreteness and session (self-report vs. instructed) in Experiment 1. Error bars represent 95% between-subjects confidence intervals.



1.2 Experiment 2

Figure S4

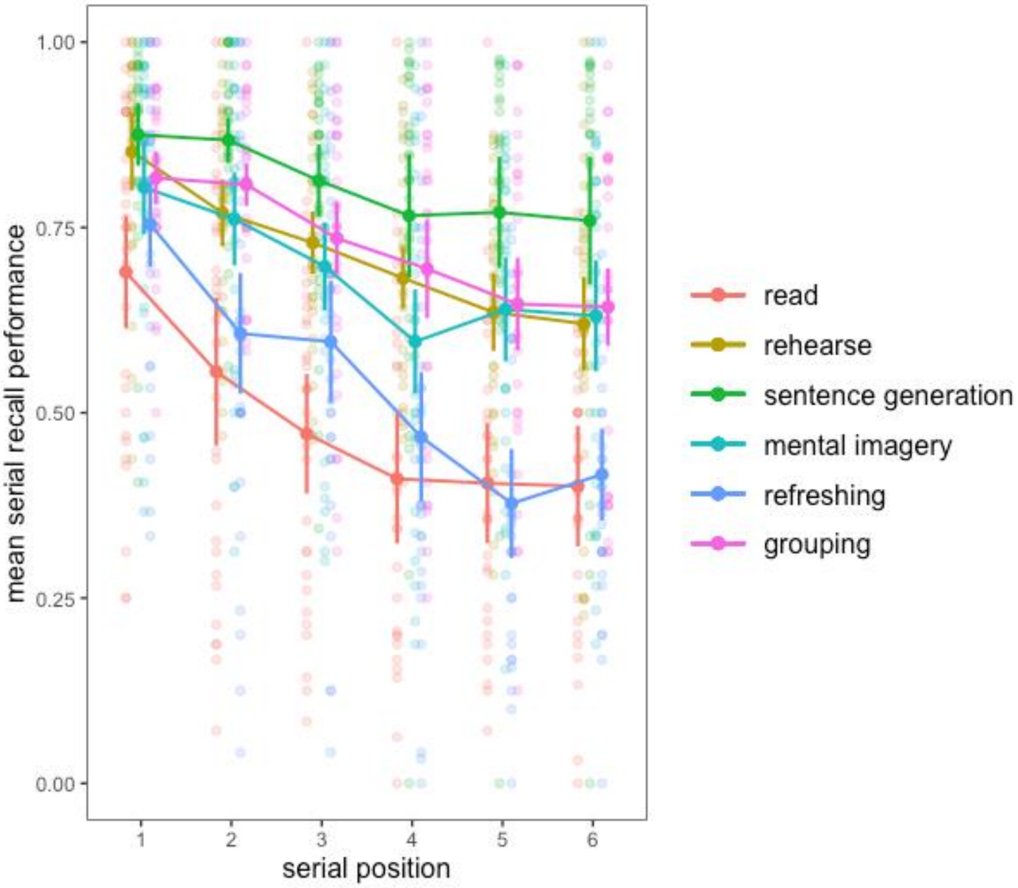
Immediate serial recall performance across serial position, strategies and session (self-report vs. instructed in Experiment 2. Error bars represent 95% between-subjects confidence intervals.



The generalized mixed effects model on data of Experiment 2 revealed strong evidence *against* the three-way interaction effect of serial position x session x strategy ($BF_{01} = 1.61171 \times 10^{5}$). There was evidence for the two-way interaction of serial position by strategy (BF_{10}= 3.91), with the differences in immediate serial recall performance across the different strategies being credible for all serial positions except for the first and the last (SP1 BF_{01}= 36.76; SP6 BF_{01}= 27.27; see **Figure S5**). There was anecdotal evidence as well for the interaction of session by strategy [BF_{10}= 2.67]). There was evidence against the serial position by session interaction [BF_{01}= 5.29]. Finally, all main effects were credibly supported by the data (strategy [BF_{10}= inf]; session [BF_{10}= 1.02 x 10^{207}]; serial position [BF_{10}= 5.19 x 10^{137}]).

Figure S5

Immediate Serial Recall performance in Experiment 2 for each serial position and strategy (averaged over sessions). Grey dots represent the data of individual subjects. Error bars show the within subject confidence interval.



2. Strategy switches across trials

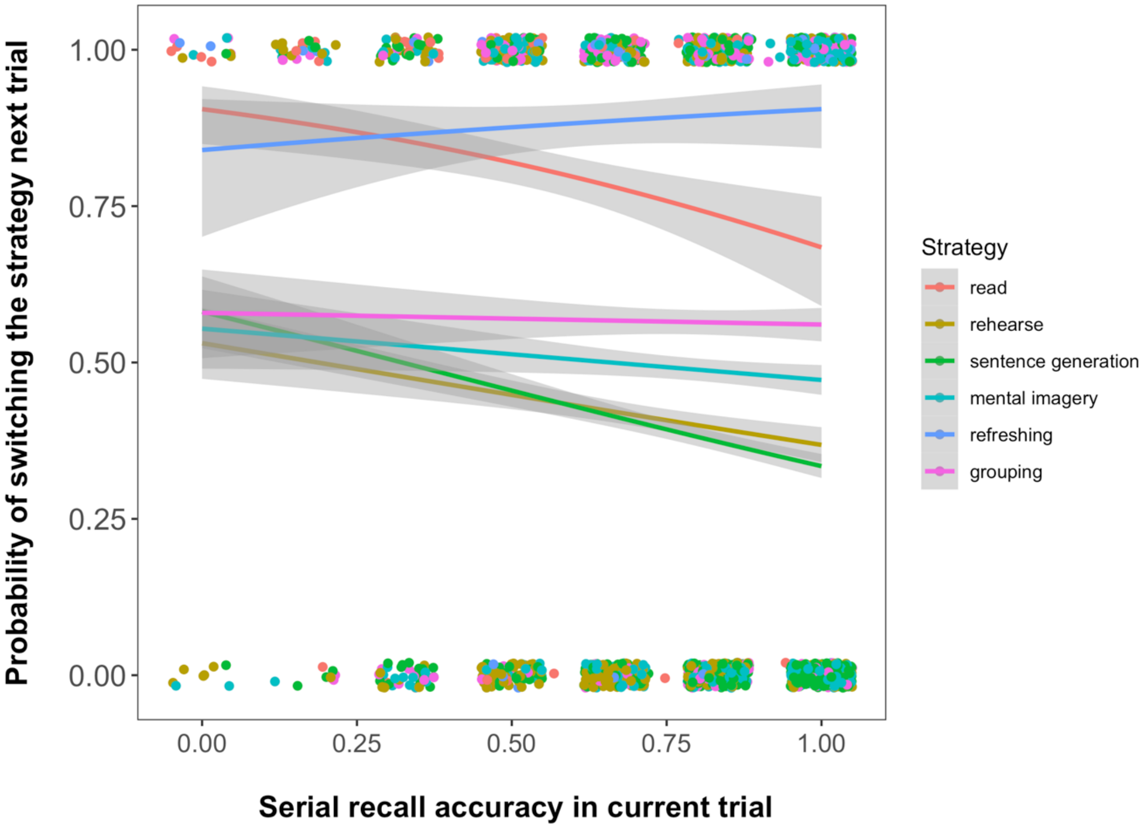
To assess whether subjects more commonly switch the strategy after showing worse recall performance in the preceding trial, we analyzed the trial-wise occurrence of a switch to another strategy in the following trial (t+1; yes vs. no), predicted by a logistic regression mixed model of the chosen strategy and the serial recall accuracy in the current trial (t).

2.1 Experiment 1

The probability of switching the strategy the next trial based on the serial recall accuracy in the current trial in shown in **Figure S6**. The analysis revealed that there was no evidence for or against an interaction of strategy with performance (BF_{01} = 1.25), and there was evidence against a main effect of serial recall accuracy (BF_{01}= 3.87).

Figure S6

The probability of switching the strategy the next trial based on the serial recall accuracy in the current trial across the strategies of the current trials in Experiment 1. For instance, the red line shows that the probability to switch from using reading in trial 1 to another strategy in trial 2 descriptively decreases the higher the serial recall accuracy was in trial 1.

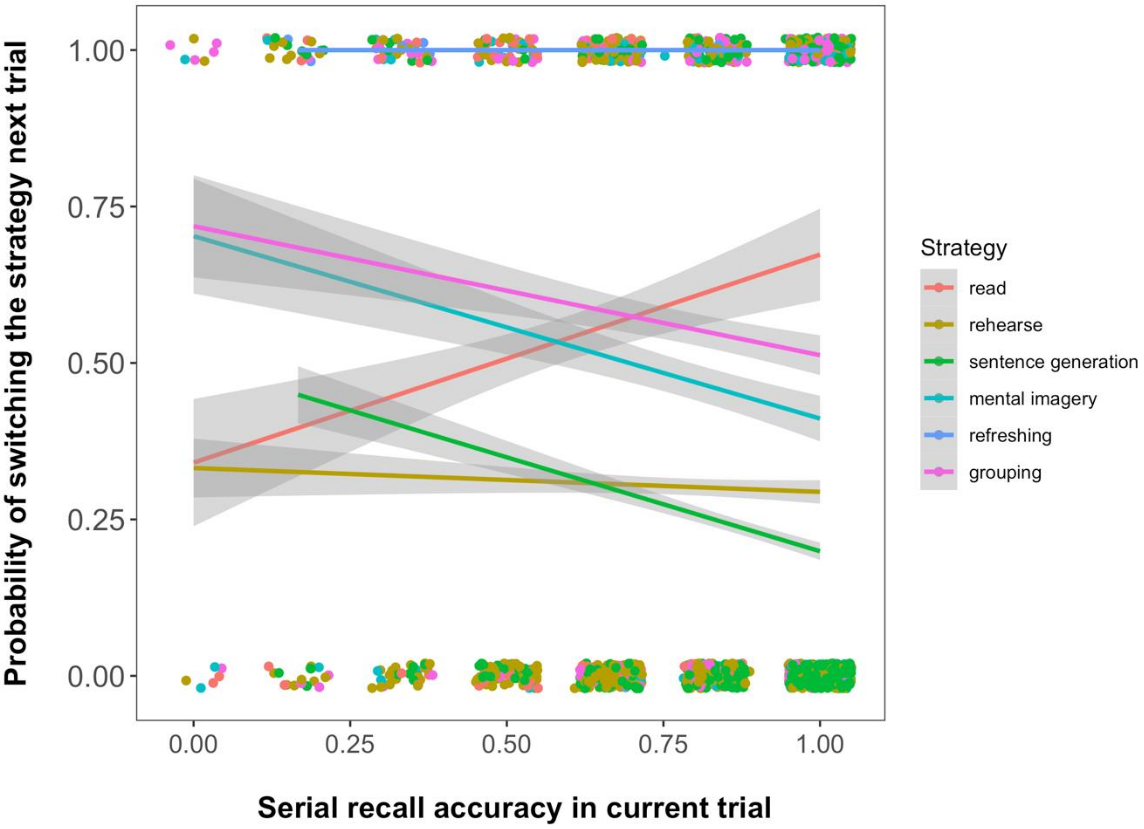


2.2 Experiment 2

The probability of switching the strategy the next trial based on the serial recall accuracy in the current trial in shown in **Figure S7**. The analysis revealed that there was no evidence for or against an interaction of strategy with performance (BF_{01} = 1.25), and as in Experiment 1, there was no evidence for a main effect of serial recall accuracy (BF_{10}= 0.73).

Figure S7

The probability of switching the strategy the next trial based on the serial recall accuracy in the current trial across the strategies in Experiment 1.



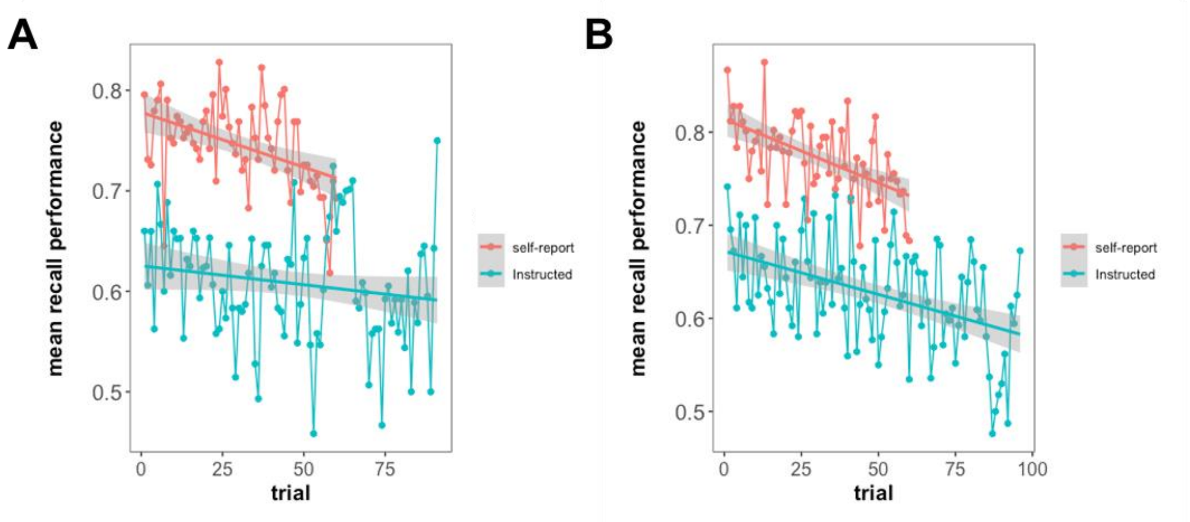
3. Effect of trial number and session on serial recall accuracy

To rule out fatigue effects within a session – especially as the second session included more trials than the first – we analyzed performance across trials (see **Figure S8**). The analysis of data in both experiments yielded evidence *against* an interaction effect of session with *trial number* (Exp.1: BF_{01} = 4.24 and Exp. 2: BF_{01} = 649.65).

In Experiment 2, we see a stronger drop in performance – independent of the session – across the trials. This is due to our manipulation of using a closed item set, therefore building up more and more proactive interference as the experiment progresses.

Figure S8

Immediate serial recall performance across trials in Experiment 1 (Panel A) and 2 (Panel B). The lines represent smoothed conditional means, using generalized linear models, and the grey zones represent the 95% confidence level interval for their predictions.



4. Instructed vs. self-reported strategies – comparing the same number of trials

Our Experiments entailed more trials for the instructed strategy session (E1: 90 trials and E2: 96 trials) compared to the self-report session (E1 and E2: 60 trials). In order to investigate whether that differences might have influenced our results, we re-ran all critical analyses of the main manuscript on data including only the first 60 trials of the instructed session. As seen in detail below, all the results remained the same and the number of trials had no effect on the main pattern of results of our study.

First, we examined whether there was evidence for an overall WM cost to implementing a strategy instruction by comparing WM performance in conditions in which a strategy was self-reported vs. instructed. Second, we examined the effectiveness of the different instructed strategies in Session 2 compared to each other. Third, we compared each instructed strategy (Session 2) to the mean performance in the self-report session (Session 1). This analysis replicates the comparison of instructed strategies to a “free strategy” baseline through which previous experiments have obtained evidence against a benefit of instructed strategies for WM performance.

4. 1 Experiment 1: Are There Dual-Task Costs of Implementing Instructed Strategies?

Figure S9 shows the immediate serial recall performance over strategies and sessions. Equivalently to the analysis including all 96 trials of the instructed session, the Bayesian LMEs revealed evidence for an interaction effect of session with strategy ($BF_{10} = 4.83$), as well as decisive evidence for both main effects (strategy: $BF_{10} = 6.14 \text{ x } 10^{23}$ and session: $BF_{10} = 1.78 \times 10^{4}$). Further, there was overall worse performance in the instructed ($M = 0.61$; $SD = 0.49$) compared to the self-report session ($M = 0.78$; $SD = 0.41$).

The main effect of strategy was driven by better performance for grouping, mental imagery and sentence generation compared to reading, rehearsal, and refreshing (see Table S1 for all pairwise comparisons).

To answer whether there are indeed dual-tasks costs of implementing instructed strategies, we turned to the interaction effect: Post-hoc comparisons of the interaction effect revealed that there was no difference between self-reported and instructed WM performance in case of rehearsal and sentence generation ($BF_{01} = 3.75$ and $BF_{01} = 3.51$, respectively). As apparent in **Figure *S9*** by the large error bars, reading and refreshing were very rarely self-reported, leading to more uncertain performance estimation. Still, choosing these low frequency strategies resulted in better performance than when they were instructed ($BF_{10} = 7.97$ and $BF_{10} = 12.17$, respectively). Likewise, both mental imagery and grouping resulted in better WM performance when these strategies were self-reported than instructed ($BF_{10} = 13.31$ and $BF_{10} = 5.26$, respectively).

Taken together, when controlling for the number of trials between instructed and self-reported strategy sessions, all the results remained the same, with only minor changes in the size of the evidence (i.e., the BFs).

Figure S9

Mean Immediate Serial Recall Performance Across Strategies and Session (Self-report vs. first 60 trials of Instructed) of the Data of Experiment 1. Error bars represent 95% Confidence Intervals. The Order of Strategies on the x-axis Reflects the Frequency with Which They were Reported (from Least to Most Frequent).

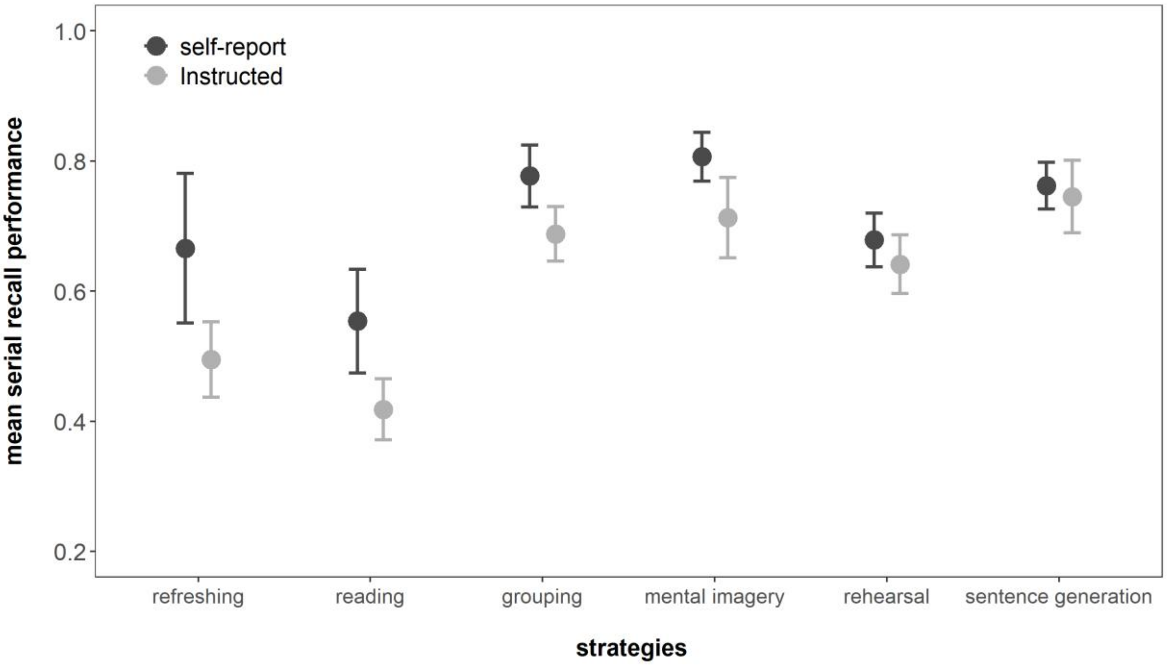


Table S1

Bayes Factors of the pairwise comparisons of *__the main effect__* *of strategy on WM performance (across conditions, but including only the first 60 trials of the instructed strategy session) of Experiment 1 and 2. Bayes Factors > 3 represent substantial evidence for better performance in the strategies listed in the columns compared to the strategy in the rows. BF < 1/3 reflect substantial evidence against a difference between both strategies*

	refreshing		rehearsal		grouping		mental imagery		sentence generation	
	E1	E2	E1	E2	E1	E2	E1	E2	E1	E2
read	0.10	0.07	8.4 × 10^{5}	3890	7.32 × 10^{9}	1.44 × 10^{7}	1.07 × 10^{12}	6513	5.16 × 10^{11}	5.49 × 10^{13}
refreshing			1.70	1155	716	2.92 × 10^{7}	2.13 × 10^{4}	210	1.03 × 10^{4}	1.07 × 10^{14}
rehearsal					9.46	0.02	1401	0.02	7765	26.96
grouping							0.72	0.04	0.02	7.10
mental imagery									1.41	9.81

4.2 Experiment 1: Which Instructed Strategy is More Beneficial?

Next, we were interested in the effect of instructed strategies on WM recall and whether our results hold when controlling for the number of trials across sessions. There was decisive evidence for a main effect of instructed strategy ($BF_{10} = 7.92 \text{ x } 10^{20}$) – equivalent to what it was before. **Table S2** presents BFs of the follow-up pairwise comparisons of the instructed strategies (only the first 60 trials) in Session 2.

These effects revealed that instructing reading or refreshing led to worse performance than all other strategies. Instructed rehearsal, grouping, and mental imagery led to similar performance. Again, these results are similar to the analysis including all trials, except for reading vs. refreshing now leading to inconclusive evidence about their difference.

Table S2

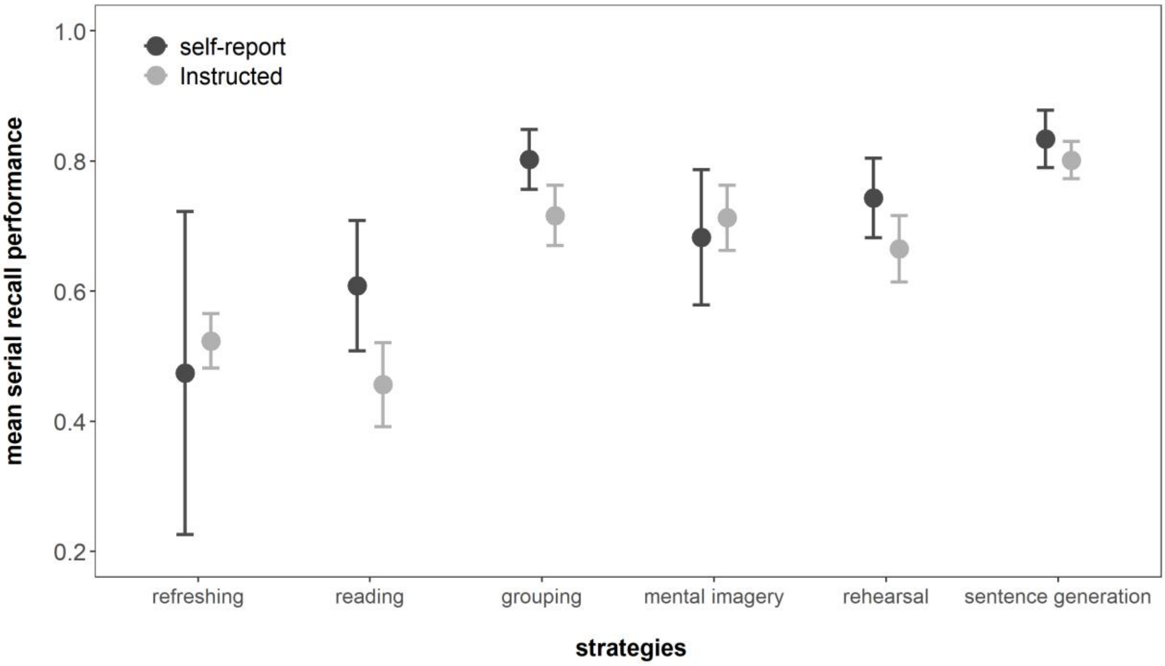
Bayes Factors of the pairwise comparisons of instructed-strategy effects on WM performance in the first 60 trials of Session 2 of Experiment 1 and 2. Bayes Factors > 3 represent substantial evidence for better performance in the strategies listed on the top compared to the ones listed on the left.

	refreshing		rehearsal		grouping		mental imagery		sentence generation	
	E1	E2	E1	E2	E1	E2	E1	E2	E1	E2
read	1.31	1.12	2.85×10^5	9847	6.77×10^7	8.48×10^5	6.50×10^6	5.51×10^6	3.31×10^7	1.54×10^{11}
refreshing			110	99.46	8283	6040	4622	1.16×10^4	1.29×10^5	4.57×10^9
rehearsal					0.58	0.45	1.06	0.82	5.19	285
grouping							0.49	0.23	0.65	5.39
mental imagery									0.37	15.93

To compare these effects of instructed strategies to a "free baseline", as done in previous studies (Bartsch et al., 2018; Bartsch & Oberauer, 2021; Souza & Oberauer, 2018, 2020), we next compared performance with each instructed strategy to the mean serial-recall performance from the self-report sessions (Session 1). None of the instructed strategies surpassed the mean serial-recall performance of the session with free strategy choice, consistent with previous literature. Instructed reading ($BF_{10} = 4.18 \times 10^7$), refreshing ($BF_{10} = 4.71 \times 10^4$), and rehearsal ($BF_{10} = 40.87$) led to worse performance than in Session 1, and instructed sentence generation and mental imagery yielded equivalent performance to the mean performance in Session 1 ($BF_{01}= 4.70$ and $BF_{01}= 3.33$, respectively). There was inconclusive evidence in case of grouping (BF_{01}= 1.42 respectively). Except for the case of mental imagery which now showed evidence against a difference (ambiguous before), there was no difference to the analysis including all the trials.

Figure S10

Mean Immediate Serial Recall Performance Across Strategies and Session (Self-report vs. first 60 trials of Instructed) of the Data of Experiment 2. Error bars represent 95% Confidence Intervals. The Order of Strategies on the x-axis Reflects the Frequency with Which They were Reported (from Least to Most Frequent).



4.3 Experiment 2: Are There Dual-Task Costs of Implementing Instructed Strategies?

Figure S10 shows the immediate serial recall performance over strategies and sessions. Equivalently to the analysis including all 96 trials of the instructed session, the Bayesian LMEs revealed anecdotal evidence for an interaction effect of session with strategy (BF_{10} = 2.44), as well as decisive evidence for both main effects (strategy: BF_{10} = 1.34 x 10^{100} and session: BF_{10} = 4.25 × 10^5). Replicating Experiment 1, as well as the analysis including all the trials, there was overall worse performance in the instructed (*M* = 0.64; *SD* = 0.48) compared to the self-report session (*M* = 0.78; *SD* = 0.41).

The main effect of strategy in Experiment 2 was driven by better performance for grouping, mental imagery and sentence generation compared to reading and refreshing (see Table S1 for all pairwise comparisons). All these patterns were the same compared to the analysis including all the trials.

As an independent test of our hypothesis of whether there are indeed dual-tasks costs of implementing instructed strategies, we turned to the interaction effect of Experiment 2 next: post-hoc comparisons revealed that there was evidence against a recall difference between self-reported vs. instructed mental imagery and refreshing (BF_{10}= 3.06 and BF_{10}= 2.66, respectively). Instruction resulted in worse performance than self-report for reading (BF_{10}= 7.53) and grouping (BF_{10}= 2.15), whereas evidence was ambiguous for rehearsal

(BF_{10}= 1.54) and sentence generation (BF10= 0.49). With this analysis it becomes clear that including fewer trials results in more ambiguous Bayes Factors -as can be expected. The general pattern of findings remains.

4.4 Experiment 2: Which Instructed Strategy is More Beneficial?

Next, we were interested in the effect of instructed strategies on WM recall. Equivalent to Experiment 1, there was decisive evidence for a main effect of instructed strategy (BF_{10} = 6.37 x 10^{23}). **Table S*2*** presents BFs of the follow-up pairwise comparisons of the instructed strategies in Session 2. These effects fully replicated the analyses including all the trials.

Taken together, when controlling for the number of trials between instructed and self-reported strategy sessions, all the results remained the same, with only minor changes in the size of the evidence (i.e., the BFs).

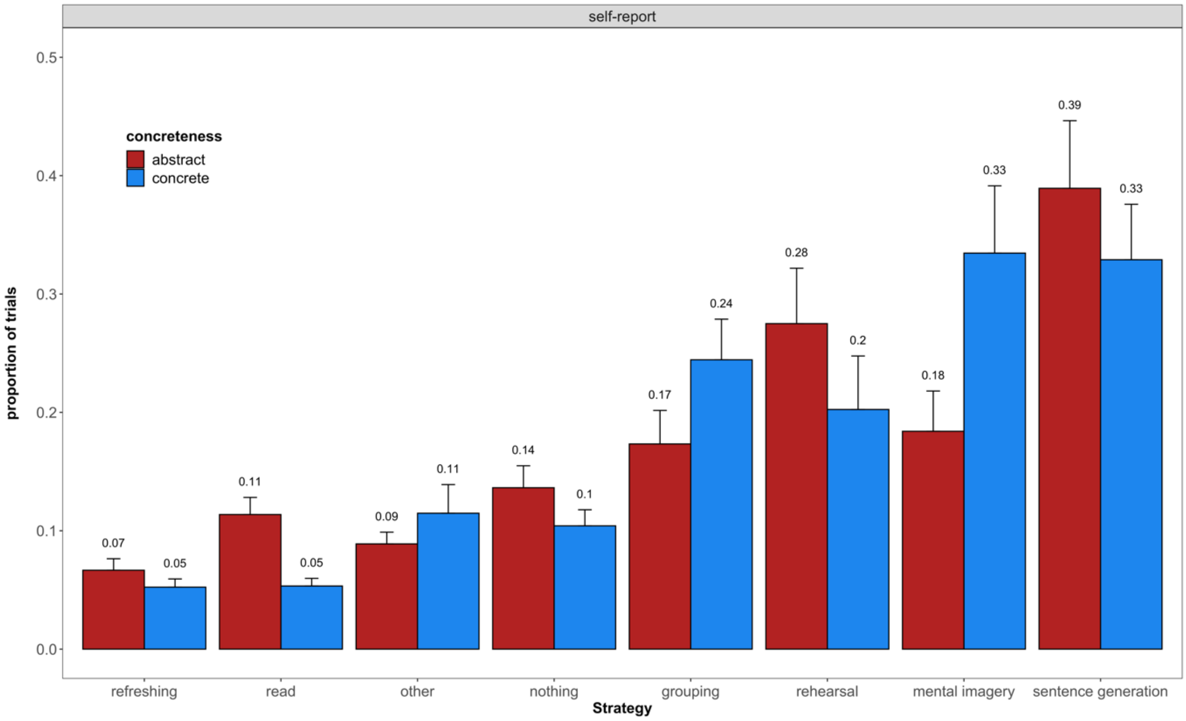
To compare these effects of instructed strategies to a "free baseline", as done in previous studies (Bartsch et al., 2018; Bartsch & Oberauer, 2021; Souza & Oberauer, 2018, 2020), we next compared performance with each instructed strategy to the mean serial-recall performance from the self-report sessions (Session 1). Here, instructed reading (BF_{10} = 1.69× 10^7), refreshing (BF_{10} = 1.42 × 10^7), rehearsal (BF_{10}= 59.91), produced worse performance compared to Session 1. Whereas instructed mental imagery (BF_{10}= 1.54), grouping (BF_{10}= 1.506495), and sentence generation yielded ambiguous evidence with regards to the comparison to the mean performance in Session 1 (BF_{01}= 1.73).

With this analysis it becomes clear that including fewer trials results in more ambiguous Bayes Factors - as can be expected. The general pattern of findings remains.

5. Effect of concreteness on probability of choosing each strategy

Figure S11

The mean proportion of trials in which each strategy was reported depending on word concreteness in Experiment 1. This variable was calculated by first computing the mean proportion of trials in which each strategy was reported by each subject, within each concreteness and then averaging across all subjects.



We analyzed the proportion of trials for which a given strategy was chosen data using the BayesFactor package (*lmbf*). The dependent variable was the probability to choose a strategy across all trials. The data are shown in Figure S9.

The linear mixed effects model on the data of Experiment 1 revealed strong evidence for the two -way interaction of concreteness by strategy (BF_{10}= 1.61 × 10^7) on the probability to choosing each strategy, Follow-up comparisons revealed that there was a credible effect of concreteness to choosing mental imagery (BF_{10} = 8.32) and grouping (BF_{10} = 5.23), with higher probabilities for concrete trials.
The opposite was the case for rehearsal and read: those were more likely to be chosen in abstract than concrete trials (BF_{10} = 8.25 and BF_{10} = 10.13, respectively). Evidence was ambiguous for sentence generation and refreshing (BF_{10} = 1.05 and BF_{10} = 1.68, respectively).

Finally, the main effect of strategy was credibly supported by the data (BF_{10}= 6.93 x 10^{26}), and there was evidence *against* an overall effect of concreteness (BF_{01}= 2.12x10^{-5}).

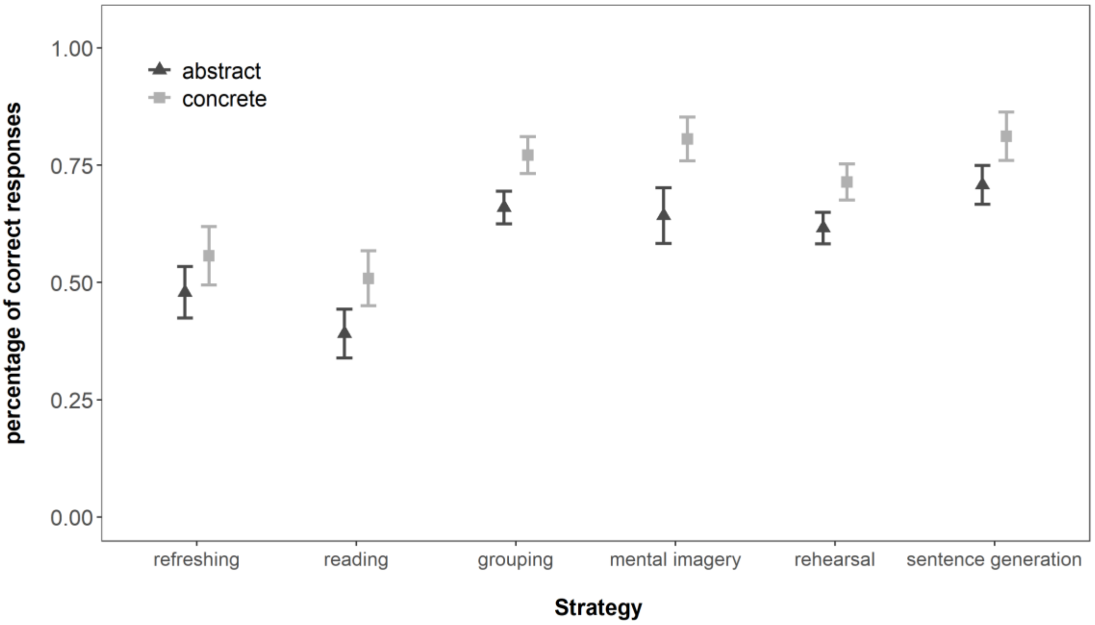
6. Does Concreteness Interact with the Effectiveness of Strategies?

Working Memory

One could expect that the concreteness of the to-be-remembered words affected how beneficial a certain strategy would be for immediate memory performance. We have seen above that people adapted their choice of strategy based on the word concreteness. Do these strategies aid more or less in recalling concrete vs. abstract words? And critically, is there evidence for an interaction, in the direction that the concreteness differentially affects some strategies but not others? To investigate this, we entered concreteness as a factor to the BGLMM. Although there was clear evidence for a main effect of concreteness ($BF_{10} = 1.8 \times 10^{13}$, see also **Figure S12**), concreteness did not credibly enter any of the interactions (Concreteness x Strategy: $BF_{10} = 0.06$, Concreteness x Session: $BF_{10} = 0.14$, Three-way interaction: $BF_{10} = 0.07$). These results indicate that abstract words are in general more difficult to recall correctly in an immediate memory test, yet that this holds independent of which strategy was employed on the memoranda, and independently of whether these strategies were self-chosen or instructed.

Figure S12

Mean Immediate Serial Recall Performance Across Strategies and Concreteness (Abstract vs. Concrete). Error bars represent 95% Confidence Intervals. The Order of Strategies on the x-axis Reflects the Frequency with Which They were Reported (from Least to Most Frequent).

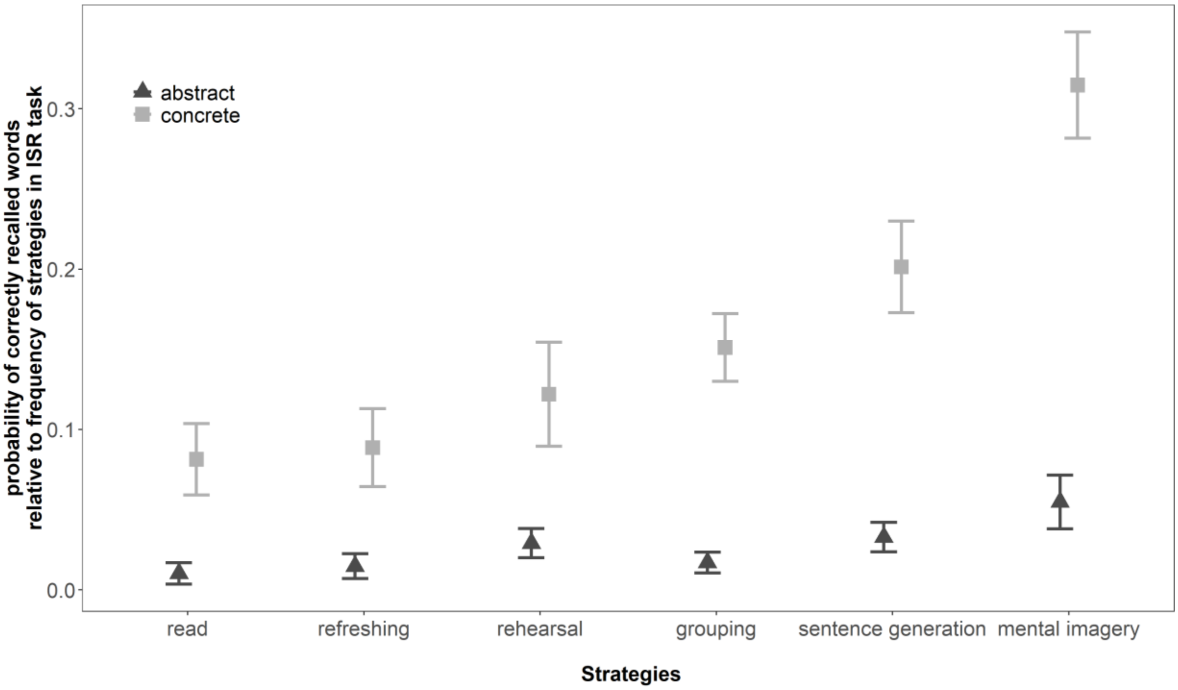


Delayed Memory

We expected that the concreteness of the to-be-remembered words affected how beneficial some of the strategies – in particular, mental imagery -- would be for delayed memory performance. **Figure S13** shows that this was the case: Concrete words benefited more from imagery, and also from sentence generation and grouping. The BGLMM provided evidence for a main effect of concreteness (BF_{10} = 7.34 x 10^{19}), with better delayed memory performance for concrete words, as well as evidence for the interaction with Strategy (BF_{10} = 3781). Follow-up analyses revealed that there were credible effects of concreteness for all strategies, which were larger for the elaborative strategies (reading: BF_{10} = 11.07, refreshing: BF_{10} = 15.06, rehearsal: BF_{10} = 29.87, grouping: BF_{10} = 1461, sentence generation: BF_{10} = 1754, and mental imagery: BF_{10} = 2.70 × 10^5).

Figure S13

Mean Probability of Correctly Recalling Items in the Delayed Free Recall Test as a Function of Strategy and Concreteness (abstract vs. concrete) relative to the Frequency of each Strategy in Experiment 1 – with an Open Pool of Items. Error bars Represent the Standard Error of the Mean.



7. Strategy Repertoire

We analyzed the mean number of strategies reported at least once by participants. The results for Exp 1 showed that participants engaged on average in 5 out of 6 strategies at least once. In Experiment 2 they engaged in 4.23 strategies at least once. So yes indeed, participants showed a large repertoire. The tables below show the number of participants using a strategy at least once and those they used at least twice. The majority of participants (>24) in Exp. 1 used a repertoire consisting of rehearsal, mental imagery, grouping and sentence generation at least twice. In Exp 2, this repertoire was used by less (> 18) participants, but still the majority. This is similar to other self-report studies, in which participants seem to report a large variety of strategy use across trials (people seem to try them at least once), see for example Morisson et al. (2016), AuBuchon and Wagner (2023). This variability in the number of used strategies appears even when the self-reports are retrospective (as in AuBuchon and Wagner, 2023) for which the strategy report could not influence the choice of applied strategies. We can only speculate whether participants felt pressured to try out each of the strategies due to our instructions, although we made sure that our instructions made it clear, that they do not have to engage in all of them, but instead only report what they do naturally. In an effort to quantify this the table below not only shows the number of participants who engaged in each strategy at least once but also the number of participants who did it at least twice – meaning using it more than in a single "try out" trial. For Exp 1 and Exp 2, it seems that refreshing and reading was rarely part of a long-lasting repertoire.

Experiment 1		
Strategy	**Number of participants engaged at least once**	**Number of participants engaged at least twice**
rehearsal	29	28
mental imagery	28	26
grouping	28	26
sentence generation	26	25
reading	21	16
refreshing	18	9

Experiment 2		
Strategy	**Number of participants engaged at least once**	**Number of participants engaged at least twice**
rehearsal	30	27
sentence generation	27	23
mental imagery	22	18

grouping	21	18
reading	19	14
refreshing	8	4