Journal of Memory and Language 105 (2019) 1–18

Contents lists available at ScienceDirect

Journal of Memory and Language

journal homepage: www.elsevier.com/locate/jml

Revisiting the attentional demands of rehearsal in working-memory tasks

Mirko Thalmann, Alessandra S. Souza, Klaus Oberauer*

Department of Psychology, University of Zurich, Switzerland

ARTICLE INFO

Keywords:
Working memory
Central attention
Articulatory rehearsal
Elaboration
Attentional refreshing

ABSTRACT

There is a recent surge of interest in maintenance processes in working memory, such as articulatory rehearsal, elaboration, and attentional refreshing. Yet, we know little about the central attentional demand of these processes. It has been assumed that articulatory rehearsal does not require central attention at all (Vergauwe, Camos, & Barrouillet, 2014), being in essence a cost-free strategy. In contrast, elaboration and attentional refreshing are assumed to incur large and continuous costs on central attention. We tested these assumptions in three experiments in which participants were presented with a varying number of words to rehearse. Participants were instructed to rehearse the words aloud, or to elaborate them by creating interactive images. Attentional refreshing was examined in a condition in which words were to be maintained during articulatory suppression. During retention participants carried out a series of choice reaction tasks, which were used to measure central attentional demands of the maintenance strategies. Articulatory rehearsal had costs on processing RTs that lasted for 10 s. Maintenance of words during articulatory suppression did not yield persistent costs on central attention, implying that participants did not continuously refresh the words. Finally, the current results cast doubt on the idea that elaboration requires central attention for an extended period of time.

All experimental scripts and data sets reported here are available online (https://osf.io/69p8j/).

Introduction

The ability to keep information in mind for carrying out complex tasks relies on working memory (WM). The capacity of WM to keep representations accessible is severely limited. In an attempt to bypass their severe WM capacity limits, people often engage in maintenance processes, which can be generically subsumed under the term rehearsal. Here, we distinguish between three types of rehearsal that have been proposed in the WM literature: articulatory rehearsal, elaboration, and attentional refreshing. The aim of the present study is to investigate the attentional demands of these three forms of rehearsal.

Three forms of rehearsal

Articulatory rehearsal is the overt or covert speaking of verbal material to oneself (e.g., Baddeley, 1996). Everyday observation, studies using overt-rehearsal protocols (Rundus & Atkinson, 1970; Tan & Ward, 2008), and laboratory self-report measures (Bailey, Dunlosky, & Kane, 2011; Dunlosky & Kane, 2007) show that people can engage in articulatory rehearsal, and often do so spontaneously. Accordingly, many models of WM attribute to articulatory rehearsal an important function for the retention of information over the short term (e.g., Baddeley, 1986; Camos, Lagner, & Barrouillet, 2009). Whether articulatory rehearsal really helps short-term retention is a matter of ongoing debate (Lewandowsky & Oberauer, 2015). In contrast, the role of articulatory rehearsal in establishing durable representations over the long term has been found to be minimal at best. Greene (1987) reviewed several studies that investigated the role of articulatory rehearsal for long-term retention, and concluded that articulatory rehearsal benefits recognition but not recall, in line with the assumption that rehearsal strengthens item memory but hardly memory for relations.

Although overt and covert rehearsal differ in terms of motor processing (e.g., overt rehearsal requires an overt articulation and a preparation thereof, but covert rehearsal does not), their effects on memory are the same. In an immediate serial recall task, Tan and Ward (2008) observed equivalent overall performance, as well as similar serial position curves and effects of presentation rate, when participants completed the task under an overt rehearsal or a silent instruction. Research from our lab (Souza & Oberauer, unpublished) replicated Tan and Ward's findings and extended them to conditions in which presentation of memoranda was interleaved with distraction (i.e., a complex-span task) in line with the idea that the effects of covert and overt articulatory rehearsal on memory are the same.

Elaboration consists of enriching to-be-remembered items with

* Corresponding author at: Department of Psychology, Cognitive Psychology Unit, University of Zürich, Binzmühlestrasse 14/22, 8050 Zurich, Switzerland.
E-mail address: k.oberauer@psychologie.uzh.ch (K. Oberauer).

https://doi.org/10.1016/j.jml.2018.10.005
Received 20 November 2017; Received in revised form 29 September 2018
0749-596X/ © 2018 Published by Elsevier Inc.

already existing representations in long-term memory (LTM; Craik & Tulving, 1975). These representations can be of any type, for example semantic or visual. Elaboration also includes the build-up of new relations between the items to be remembered, such as connecting the words of a list to be remembered in a sequence (Craik & Tulving, 1975; Greene, 1987). Laboratory self-report measures (Bailey et al., 2011; Dunlosky & Kane, 2007) show that in about one third of the trials participants spontaneously engage in elaboration during a WM task using words as memoranda. Elaboration has a beneficial effect on remembering items over the long term (Craik & Tulving, 1975), but whether it has a beneficial effect over the short term is currently unclear. Recent experimental evidence suggests elaboration does not improve WM (Bartsch, Singmann, & Oberauer, 2018), but there is correlational evidence in support of a beneficial effect of elaboration from studies asking trial-by-trial reports of rehearsal strategies (Bailey, Dunlosky, & Kane, 2008).

Lastly, *attentional refreshing* involves briefly thinking of a representation in WM, hence bringing this representation to the focus of attention, and thereby extending and/or augmenting its accessibility (Raye, Johnson, Mitchell, Greene, & Johnson, 2007). Refreshing is assumed to be a sequential process, in which only one item can be refreshed at any point in time (but see Portrat & Lemaire, 2014), and the time required for refreshing one item has been estimated to around 35–50 ms (Vergauwe, Camos, & Barrouillet, 2014; Vergauwe & Cowan, 2014). The role of refreshing for maintenance in WM has been mainly inferred from the observation of better memory performance when participants have some free time in between the memoranda or processing episodes in WM tasks (Barrouillet, Bernardin, & Camos, 2004; Camos et al., 2009). Only a few studies actually manipulated attentional refreshing as an independent variable in a WM task and examined its effects on short-term memory performance (Bartsch et al., 2018; Souza & Oberauer, 2017; Souza, Rerko, & Oberauer, 2015; Souza, Vergauwe, & Oberauer, 2018; Vergauwe & Langerock, 2017). For example, Souza et al. (2015) investigated refreshing by asking participants to attend to individual visual WM representations during the retention interval. That study showed that the number of refreshing attempts positively affected WM. In contrast, the study by Bartsch et al. (2018) did not find a beneficial effect of refreshing on WM.

Attentional costs of rehearsal

Here, we are interested in the degree to which articulatory rehearsal, elaboration, and refreshing demand central attention. Central attention is a capacity-limited processing mechanism that has been characterized as a bottleneck that enforces serial processing (Pashler, 1994), or a processing resource that constrains the speed of parallel processes (Tombu & Jolicoeur, 2003). Dual-task studies have shown that central attention is involved in response selection and retrieval from LTM (Johnston, McCann, & Remington, 1995). Here we ask to what extent the three forms of rehearsal rely on central attention.

Knowledge about the central attentional demands of different forms of rehearsal plays an important role in determining how they can be applied in a WM task. Moreover, an understanding of how much central attention is needed for a specific rehearsal strategy is essential in predicting how these strategies can be combined. For example, the time-based resource sharing (TBRS) model of WM assumes that articulatory rehearsal and refreshing can be applied simultaneously in WM tasks (Camos, 2015; Camos & Barrouillet, 2014; Camos et al., 2009; Mora & Camos, 2015; Mora & Camos, 2013). The assumption within the TBRS model is that articulatory rehearsal does not require central attention at all, being in essence a cost-free strategy. In contrast, refreshing is assumed to depend on central attention. Therefore, carrying out other attentionally demanding tasks either postpones refreshing, or execution of these tasks is postponed because refreshing is taking place (hence showing a trade-off). If, however, articulatory rehearsal is not attentionally demanding, it can be applied simultaneously with refreshing with no trade-offs.

The assumption that articulatory rehearsal does not demand (or demands very little) central attention rests on the findings of two early studies (Guttentag, 1984; Naveh-Benjamin & Jonides, 1984). In the following, we will briefly summarize these studies and discuss why conclusions from them regarding the central attentional demands of articulatory rehearsal are problematic.

Guttentag (1984) examined the trade-offs between carrying out overt articulatory rehearsal of a memory list simultaneously with a finger tapping task. Specifically, he compared how many times per minute children from different age groups were able to tap with their forefinger in a single-task condition, and in a dual-task condition simultaneously requiring articulatory rehearsal. A core result of his experiments was that articulatory rehearsal led to severe costs on the tapping task. The dual-task costs correlated negatively with the age of the children. The interpretation of Guttentag was that articulatory rehearsal becomes more and more automatized with age, which is why his paper is often cited in support of articulatory rehearsal being non-demanding in adults. However, the dual-task cost of articulatory rehearsal in the oldest children (mean age = 11.5 years) was still around 15%, clearly speaking against complete automatization of articulatory rehearsal.

Naveh-Benjamin and Jonides (1984) used a sophisticated design to examine several questions concerning articulatory rehearsal and elaboration. Participants were given three two-digit numbers on every trial to remember for immediate recall (pre-load task). Next, two words were presented to be rehearsed during retention of the numbers. One group was instructed to rehearse synchronously with a metronome (articulatory rehearsal group); the other group was instructed to elaborate the words independently of the metronome beat (elaboration group). Whereas the elaboration group was informed about a delayed recognition test of the words at the end of the experiment, the articulatory rehearsal group was not. A dot appeared on the screen either 0.850 s, 4.675 s, or 12.325 s after presentation of the two words (hence while participants were rehearsing or elaborating the two words). Participants were instructed to press a button as soon as they detected the dot on the screen. After that, recall of the three two-digit numbers was required, ending the trial. The most important result for the present examination is that the reaction time (RT) to detect the dot was on average 22 ms faster when it appeared 12.325 s after start of rehearsal than when it appeared 4.675 s thereafter. There was no such decrease in RTs over time for the elaboration group. Based on these results, the authors concluded that articulatory rehearsal – but not elaboration – can be executed without the requirement of attention after an initial set-up stage.

There are several reasons why these two studies do not allow clear statements about the central attentional demands of rehearsal to be made. First, both studies assessed dual-task costs on tasks that do not require response selection. Work with the psychological refractory period (PRP) effect has shown that simple RT tasks such as the ones employed in these studies do not engage the central attentional bottleneck (Pashler, 1994). Based on this research, it is difficult to argue that what was measured by the RT tasks in Guttentag (1984) and Naveh-Benjamin and Jonides (1984) was central attention. Nevertheless, the dual-task cost of rehearsal on tapping in Guttentag's study was substantially larger than zero across all age groups, contrary to the assumption that articulatory rehearsal is cost-free.

Second, in the experiment by Naveh-Benjamin and Jonides there was no control condition to compare the effects of rehearsal to. Therefore, it is unclear whether dual-task costs remained for articulatory rehearsal even after 12 s or more. Accordingly, this study may have led to an underestimation of the attentional cost of articulatory rehearsal. Third, the requirement to rehearse in synchrony with a metronome could have compromised the measurement of attentional costs of articulatory rehearsal in Naveh-Benjamin and Jonides' (1984) study. The longer RTs to dots presented earlier during the rehearsal time could

reflect some initial attentional cost of adapting the pace of articulatory rehearsal to the metronome. Similarly, the fact that RTs in the elaboration group did not decrease over the processing phase may be because participants did not have to align their rehearsal with the metronome. To summarize, the evidence that articulatory rehearsal does not require central attention based on these two studies is weak, at best.

The present study

The concerns raised above indicate that it is time to revisit the attentional demands of rehearsal. In the present experiments we made an effort to obtain a proper estimate of the central attentional requirements associated with articulatory rehearsal and elaboration while avoiding the pitfalls previously mentioned. In addition, we assessed the attentional demands of a condition in which participants were required to perform articulatory suppression (AS) during maintenance of a memory list. Vergauwe et al. (2014) have argued that blocking the use of articulatory rehearsal with AS prompts participants to resort to refreshing of the memoranda. Our goal was to evaluate whether attentional costs in this condition are consistent with the idea that participants spontaneously engage in refreshing.

Experiment 1 examined the central attentional demands of articulatory rehearsal and elaboration in a design closely modeled after Naveh-Benjamin and Jonides (1984). The benefit of this design is that participants executing articulatory rehearsal do not anticipate the memory test for the rehearsed material, and therefore have no incentive to engage in additional processing of the rehearsed material. In this way, central attentional requirements of articulatory rehearsal can be measured while controlling for any additional type of rehearsal that participants may spontaneously use. In Experiment 2 we zoomed in on articulatory rehearsal because of the importance of the assumption that it is a cost-free strategy in the TBRS model (e.g., Camos et al., 2009). We investigated the attentional costs of articulatory rehearsal with a paradigm previously used by proponents of the TBRS theory to investigate attentional costs of maintenance processes (Vergauwe et al., 2014). In Experiment 3, we again compared articulatory rehearsal and elaboration with each other, and we assessed whether people spontaneously engage in refreshing under AS.

In all three experiments we used a choice RT (CRT) task to probe the attentional demand because this task requires response selection, and hence requires central attention (Pashler, 1994). Furthermore, we applied the overt rehearsal methodology (Rundus & Atkinson, 1970) to check for compliance with the articulatory rehearsal instruction. Without requiring overt responses, it is difficult to make inferences about the use of articulatory rehearsal. Moreover, we tested for LTM at the end of the experiments to assess whether participants engaged in elaboration.

To foreshadow our results, articulatory rehearsal yielded dual-task costs on CRT, which increased with the number of items to be rehearsed. The costs of both articulatory rehearsal and elaboration vanished after about 5 s when no WM test was forthcoming. When a WM test was expected, these costs persisted until the end of the processing period (10 s). In contrast, we did not observe persistent dual-task costs on CRTs in the AS condition supposed to engender refreshing (cf. Vergauwe et al., 2014).

Experiment 1

In Experiment 1 we gauged the attentional costs of articulatory rehearsal and elaboration in an adapted version of Naveh-Benjamin and Jonides (1984) experiment. As in their seminal study, we assigned participants to one of two groups: Articulatory Rehearsal or Elaboration. In each trial, participants were presented with 2 words, which they had to read aloud. Thereafter, the Articulatory Rehearsal group was instructed to continuously rehearse the words aloud, whereas the Elaboration group was instructed to create vivid and interactive mental images of the meaning of these words. To control for compliance with the instructions, we recorded the speech of participants in the Articulatory Rehearsal group. Moreover, we presented a delayed recognition test of the memoranda at the end of the experiment. In keeping with Naveh-Benjamin and Jonides (1984), the Elaboration group, but not the Articulatory Rehearsal group, was informed at the beginning of the experiment about the final memory test. This information served as a motivation to carry out the elaboration strategy. Better delayed recognition performance of the Elaboration group in comparison to the Articulatory Rehearsal group is expected if they complied with the instructions.

To assess the time course and magnitude of the attentional costs of articulatory rehearsal and elaboration, offset of the words was followed by a 10-s processing phase in which dots were presented at unpredictable, irregular periods, and participants made speeded judgments about their locations. Our testing of the attentional costs of rehearsal deviates from Naveh-Benjamin and Jonides (1984) in five regards. First, a CRT was used instead of a simple RT (SRT) to measure central attentional demand. Second, a control condition was included in which no words had to be rehearsed at all, hence serving as a single-task baseline for the CRT task within each group. Third, the processing phase in every trial lasted 10 s and was divided into five segments. Within each segment, a CRT stimulus was shown with a fixed probability. Fourth, we omitted the pre-load of three two-digit numbers used by Naveh-Benjamin and Jonides (1984) to avoid any possible confounds between the costs of rehearsing the words and any other potential rehearsal participants would attempt to maintain the pre-load memoranda. Fifth, we also dropped the requirement for participants to articulate in synchrony with a metronome. The requirement to align a speech output with a metronome may need central attention, which would have compromised the attempt to measure the central attentional demands of articulatory rehearsal alone.

To summarize, Experiment 1 had three independent variables: rehearsal instruction (articulatory rehearsal vs. elaboration, varied between-participants), set size (two vs. zero words to be rehearsed, varied within participants), and time segment within which a processing stimulus was presented (1–5, varied within-participants).

Method

Participants

Fifty-four university students (34 women; $M = 26$ years old, $SD = 5.4$ years) were randomly assigned to one of two groups: Articulatory Rehearsal or Elaboration. All participants in the experiments reported here were compensated with partial course credit or 15 Swiss Francs for one 1-h session. All participants were university students and native speakers of German. They provided written informed consent and were debriefed in the end of the experiment. Moreover, participants were informed that their speech during the experiments would be recorded and inspected to control for compliance with the instructions.

Materials and procedure

All experiments reported here were programmed in MATLAB using the Psychophysics Toolbox 3 (Brainard, 1997; Pelli, 1997). Participants were tested in individual booths where they sat at a distance of approximately 50 cm from the computer screen. They wore headphones equipped with a microphone for recording of their speech.

Lists were generated randomly for each participant. First, 100 two-word lists were constructed by randomly selecting (without replacement) from a pool of 200 mono- and disyllabic German nouns. Second, forty control lists with the letter string “xxxxx” were added to this pool, resulting in 140 lists. Finally, on every trial of the rehearsal task, one list was randomly sampled without replacement from this pool.

In every trial, the two rehearsal words were presented

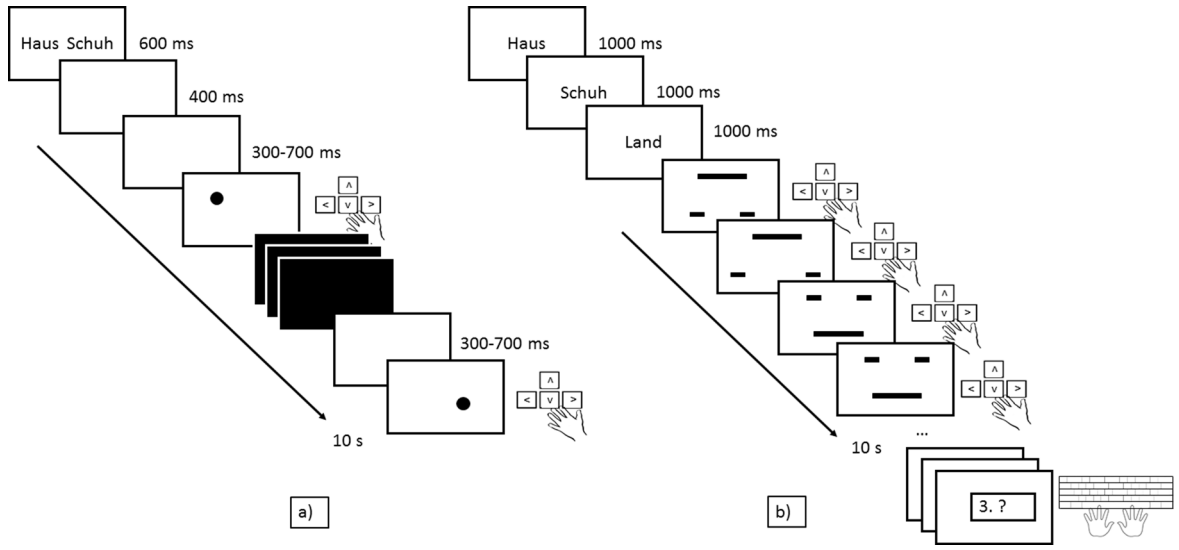


Fig. 1. Illustration of the sequence of events in a trial in Experiment 1 (panel a) and Experiments 2 and 3 (panel b). Panel a: In the beginning of each trial, two words for the rehearsal task were displayed simultaneously. Next, a 10-s period started in which participants were instructed to continuously rehearse the words while concurrently performing a dot CRT task. The dot task was divided into five 2-s segments. In each segment, a dot was presented with $p = 0.46$. The black frames represent segments of 2000 ms, in which no processing stimulus happened to be presented. Panel b: In Experiments 2 and 3, the memory words were presented sequentially. The first stimulus of the CRT was presented immediately after offset of the last word, and each CRT response was immediately followed by the next CRT stimulus until 10 s had elapsed. After that, participants were requested to attempt typed forward serial recall of the words. Stimuli are not drawn to scale.

simultaneously in the center of the screen for 600 ms followed by a blank delay of 400 ms (see Fig. 1a). Next, a 10-s interval followed in which the rehearsal task was combined with a CRT processing task (described below). Participants in the Articulatory Rehearsal group were instructed to uninterruptedly rehearse the words aloud during the processing phase. Participants in the Elaboration group were instructed to continuously elaborate the words during the processing phase for a subsequent delayed memory test. Participants were informed that in the 2-word condition their main task was to rehearse the two words according to their group's instruction, and their secondary task was to respond to the dot task. In the 0-word condition, when the "xxxxx"-string was presented on the screen, they were instructed to simply respond to the dot task (0-word condition); in the Articulatory Rehearsal group they were instructed to remain silent during that time. The 0-word condition served as a single-task baseline of responses to the dot task.

As in Naveh-Benjamin and Jonides's (1984) study, the Elaboration group was informed that elaboration is particularly helpful for remembering the words in the long-term. More specifically, participants in this group were told to: *"Create a mental image of the two words. Then, try to make an image incorporating both images, such that they interact with each other. And try continuously to change their interaction or make the image more vivid."*

In the processing task, participants had to indicate as fast and as accurately as possible whether a dot (occurring at unpredictable intervals) was shown above or below the horizontal screen midline by pressing the up or down arrow keys on the keyboard. The total duration of the processing task was 10 s, which was divided into five 2 s-segments. Within every segment, a dot (diameter = 60 pixels) was shown with $p = 0.46$. To create uncertainty about the dot's location, its horizontal position was randomly selected from a uniform distribution ranging from −250 to +250 pixels in relation to the center of the screen. The vertical position of the dot was selected such that 4/5 of the dot (48 pixels) fell either in the upper or lower part of the screen. If a dot was selected to be shown within a given 2-s time segment, its onset after the beginning of the segment was sampled from a uniform distribution between 300 and 1000 ms. The dot remained visible until an answer was given or after 500 ms of the next time segment had elapsed. An answer was counted as valid if it occurred while the dot was onscreen; otherwise, a time-out was recorded. In case an answer to a presented dot was not given in the same time segment but in the 500 ms of the next time segment, and another dot was scheduled to be presented in that time segment, the overlap of that response into the new time segment was added to the onset time of the next dot (which was again sampled from a uniform distribution between 300 and 1000 ms). The overlap was however not added to the total duration of the processing task, which did never outlast 10 s. This procedure assured that time pressure was within a reasonable range.

After completion of the 140 trials of the main experimental task, a delayed recognition test of the rehearsal words was presented for all participants. The main purpose of this test was to ensure that the Elaboration group complied with the instruction to elaborate on the presented words. For the Articulatory Rehearsal group, the delayed recognition test was a surprise. The first five and the last five rehearsal word lists were excluded from this test to control for primacy and recency effects. One noun of each of the remaining 90 lists was randomly selected as a cue. Each trial in the delayed recognition test consisted of the presentation of a cue on the left side of the screen, and a set of four candidate words (the correct word and three distractors) on the right side of the screen. One distractor was a word presented in another memory list (intrusion), and the remaining two distractors were new words selected from a pool of 180 mono- and disyllabic German nouns. Participants were instructed to select among the set of candidate words the one that was presented together with the cue during the rehearsal task. Participants responded by clicking with the left mouse button on one of the four words. The order of testing of the rehearsal words was randomly determined for each participant.

Results

Statistical framework

We used Bayesian statistics for all analyses. Bayesian statistics overcome some of the shortcomings associated with conventional null-hypothesis significance testing (Wagenmakers, 2007). In the Bayesian framework (for an introduction see Kruschke, 2011), prior knowledge about the credible values of the model parameters is expressed as probability distributions known as priors. These priors are updated in light of the data to yield posterior distributions. The posterior reflects

the knowledge about credible values of model parameters after taking into account the data.

In addition to the posteriors of credible model parameters, the relative credibility of two models can be computed with the Bayes Factor (BF). The BF is the ratio of the marginal likelihoods of two models (Rouder, Morey, Speckman, & Province, 2012). By formulating two hypotheses of interest as statistical models and assuming equal priors for the two models, the BF quantifies how many times more likely one hypothesis is than the other, given the data. For the comparison of a null hypothesis with an alternative hypothesis on a parameter of interest (e.g., the difference between two groups in a *t*-test, or an effect in an ANOVA design), the point null hypothesis is represented by a model in which the prior of the relevant parameter is set to 0. The interpretation of the BF is straightforward. A BF of 1 states that the data are ambiguous, providing no evidence favoring one model or the other. Although a BF between 1 and 3.2 states that one model is more likely than the other, it is usually "not worth more than a bare mention" (Kass & Raftery, 1995). A BF from 3.2 to 10 is regarded as substantial evidence, a BF from 10 to 100 as strong evidence, and a BF larger than 100 as decisive evidence for one model over the other.

One advantage of Bayesian statistics is that sample size does not have to be determined a priori to data collection because the chance of making a Type I error does not increase with optional stopping (Rouder, 2014). Accordingly, data collection can be continued until the evidence for the null hypothesis or the alternative hypothesis is substantial. In this and the remaining experiments in this article, an initial sample size was set to values that in our experience are sufficient to obtain robust evidence for medium-sized effects in within-subjects (and mixed) designs. Whenever evidence for the main hypotheses under consideration was ambiguous, we continued data collection until evidence was, at least, in the substantial range.

Compliance with task instructions

Six participants were excluded due to responding close to chance level in the CRT task.[1] Another participant in the Articulatory Rehearsal group was excluded because this participant mentioned after the experiment to have expected the delayed recognition test and used elaboration to better remember the words. Six additional participants were excluded, because they did not adhere to the rehearsal instructions, as assessed by the inspection of the speech recordings of the trials. For example, some participants articulated the "xxxxx"-string in the control condition instead of remaining silent, some read wrong words or only sporadically rehearsed some of the word lists. This resulted in a final sample of 41 participants, with $n = 23$ in the Elaboration group and $n = 18$ in the Articulatory Rehearsal group.

Delayed recognition

Next, we analyzed the responses in the delayed recognition test to assess whether the elaboration instruction worked. If participants in the Elaboration group complied with the instruction, we expect to observe better LTM for the rehearsal word pairs compared to the Articulatory Rehearsal group, replicating the results of Naveh-Benjamin and Jonides (1984). Responses in the delayed recognition test were classified into three categories: hits (selection of the correct alternative), intrusions (selection of the alternative with a word from another list), and false alarms (selection of one of the new words). The proportion of responses in each of these categories is shown in Table 1. Hits (log-odd transformed) in the Articulatory Rehearsal condition and the Elaboration condition were compared with a Bayesian linear regression model.

The regression model was run via JAGS (Plummer, 2003), which was accessed via the R statistical computing environment (R Development Core Team, 2015) with the *rjags* package (Plummer, Stukalov, & Denwood, 2015). We approximated the BF of the difference between conditions via the Savage-Dickey density ratio. This method obtains the BF for two nested models, such as a Null model assuming that the effect is zero, and an alternative model that allows the effect to vary freely. The first step is to obtain the posterior of the parameter of interest – here, the size of the difference between conditions in the probability of a hit – in the alternative model. The BF is then obtained by dividing the height of the posterior by the height of the prior at the parameter value assumed in the Null model (Lee & Wagenmakers, 2014; Wetzels, Raaijmakers, Jakab, & Wagenmakers, 2009). The BF was 8.42×10^6 for the Alternative hypothesis over the Null, which is decisive evidence for better delayed recognition in the Elaboration group than in the Articulatory Rehearsal group.

Table 1
Proportion of responses in each response category in the delayed recognition test in Experiment 1.

	Condition	
Response category	Articulatory rehearsal	Elaboration
Hits	0.425	0.818
Intrusions	0.293	0.095
False alarms	0.141	0.043

Note. False-Alarm rate was divided by 2, because the probability of making a False Alarm was twice that of making a Hit or an Intrusion.

Processing task

RT and accuracy to respond in the processing task served as the dependent variables. RTs were trimmed as follows. First, RTs associated with incorrect answers and time-outs were removed (15.57% of all RTs available for analysis). Second, RTs that exceeded or fell below the individual mean ± 3 standard deviations in each time segment were excluded (0.87% of the remaining RTs). The remaining RTs were averaged within each segment in each set-size condition (2 words vs. 0 words) and group (Articulatory Rehearsal vs. Elaboration). In Fig. 2, RTs (panel a) and accuracies (panel b) are plotted against segment. RTs and accuracies (log-odds transformed) were analyzed separately with $2 \times 2 \times 5$ Bayesian ANOVAs using the BayesFactor package 0.9.10-2 (Morey & Rouder, 2014), which is available in R. For all analyses with the BayesFactor package in this article we used the default combination of JZS priors (Cauchy prior on effect size, Jeffreys prior on the variance).

For each ANOVA, the winning model was selected as the one with the highest BF in comparison to the null model. Evidence for each effect included in the winning model was gauged by comparing the winning model to a model derived from it by dropping the effect in question. Conversely, evidence against an effect excluded from the winning model was assessed by adding it to the winning model.

For RTs, the winning model included fixed effects of segment, group, set size, set size x segment, and a by-subjects random intercept (BF = 6.11×10^{13} over the null model). Next, we describe the evidence for each effect of theoretical relevance. The Elaboration group responded more slowly than the Articulatory Rehearsal group (BF = 3.8). Rehearsal slowed responding in the processing task, as indicated by the slower RTs in the 2-words condition compared to the 0-word control condition (BF = 458). Moreover, RTs decreased across the five time segments (BF = 1.39×10^9). The costs of rehearsal (2-words vs. 0-words) decreased over segment (segment x set size, BF = 71.6). There was not enough evidence to support a difference between the Elaboration group and the Articulatory Rehearsal group on the decrease of RTs over segment (segment x group, BF = 1.7). The evidence was against the two-way interaction of group x set size (BF = 0.19; hence indicating that the Null was supported by a factor of 5.3). This indicates that the time cost resulting from rehearsing 2 words versus not

[1] They fell below the 99.9% quantile of a binomial distribution with mean 0.5 (i.e., less than 190 correct responses across the 325 CRT episodes presented over the experiment).

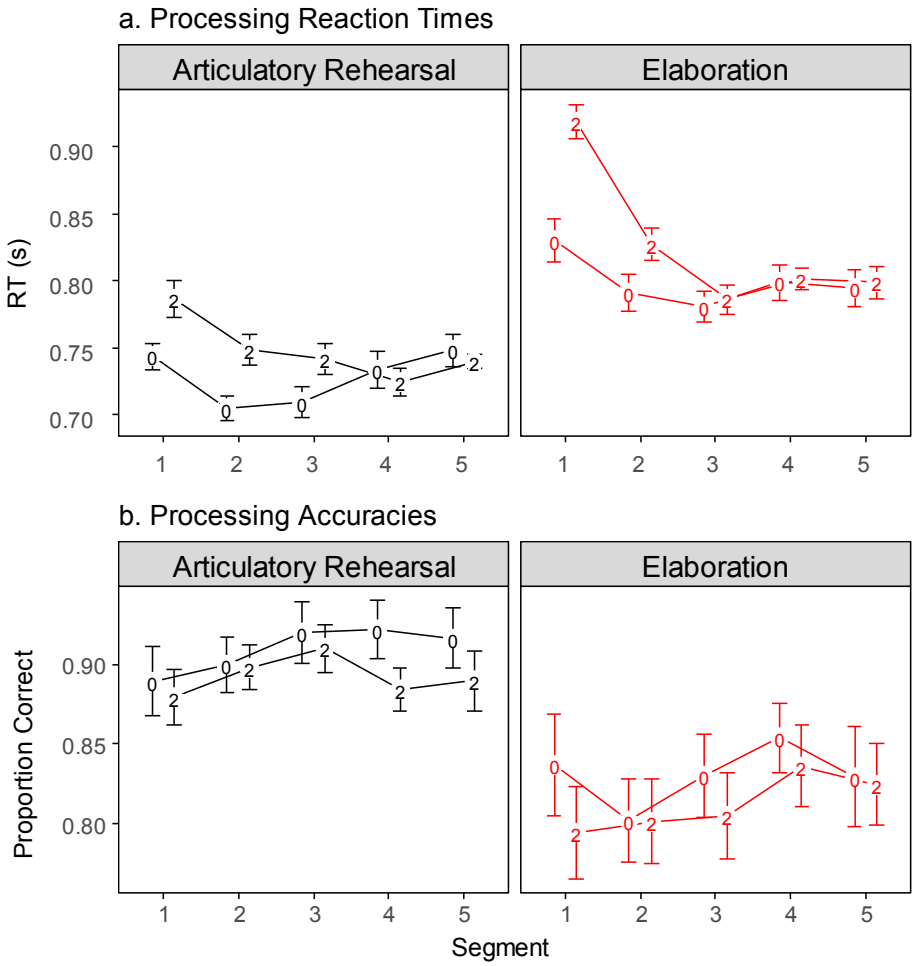


Fig. 2. RTs (panel a) and accuracies (panel b) plotted over segment for each set size and rehearsal group in Experiment 1. Error bars represent standard errors for within-subjects designs (Bakeman & Mcarthur, 1996).

rehearsing at all did not differ between the two rehearsal instructions. Moreover, there was also evidence against the three-way interaction (BF = 0.16, evidence for the Null equals 1/0.16 = 6.25), indicating that the decrease in the costs of rehearsal over segment was similar for both rehearsal groups.

For accuracies, the winning model included the fixed effects of rehearsal group and of set size, and a by-subject random intercept (BF = 11′238 over the Null). The BF in favor of rehearsal instruction was 4.7, indicating that participants in the Articulatory Rehearsal group responded more correctly, on average, than participants in the Elaboration group. The higher accuracies and faster RTs in the Articulatory Rehearsal group compared to the Elaboration group rule out a speed-accuracy trade-off to explain the group difference. The main effect of set-size yielded a BF = 2397, which implies that rehearsing 2 words credibly reduced accuracy in the processing task. The BF was 23 against the set size x segment interaction, which is strong evidence that the set-size costs on accuracy did not decrease over the processing period.

Discussion

Two observations support the conclusion that participants complied with the rehearsal instructions. First, elaboration led to far better performance in the delayed recognition test than articulatory rehearsal. Second, inspection of the recorded speech confirmed that the participants in the Articulatory Rehearsal group were articulating the words continuously.

The main findings of Experiment 1 can be summarized as follows. First, responses in the processing task were initially slower when participants had to rehearse 2 words compared to the control condition with 0 words. This shows that articulatory rehearsal and elaboration require central attention for some period of time. Second, the attentional costs of both forms of rehearsal decreased over time. Visual inspection suggests that the time costs of both types of rehearsal vanished after about five seconds; the costs on accuracy, however, were unaffected by time. Third, the magnitude of the attentional costs of rehearsal did not differ between the two rehearsal groups, showing that both articulatory rehearsal and elaboration engage central attention to a similar extent.

Using an optimized experimental design to measure central attentional demands, our results confirm the most important conclusion of Naveh-Benjamin and Jonides (1984) that articulatory rehearsal is initially attentionally demanding, but the demand declines within the first 5 s, after which the RT costs disappear. Costs on accuracy, however, remained. Different from Naveh-Benjamin and Jonides (1984), the present results suggest that the attentional cost of elaboration likewise declines over the first 5 s elaboration. Taken together, our findings suggest that both examined rehearsal strategies always incur an attentional cost. This cost is unlikely to vanish during a typical WM task such as complex span (in which memoranda and processing episodes are interleaved), because a new rehearsal process has to be initiated every time a new item is presented, and that usually happens at a rate of 1–5 s.

We observed an unexpected difference in RTs and accuracies

between the two groups: Even when there were no words to be rehearsed (set size = 0), the Elaboration group was slower and less accurate than the Articulatory Rehearsal group. This difference cannot be attributed to rehearsal of items presented on the current trial. One potential cause of this group difference is that participants in the elaboration group, expecting the final recognition test, engaged in additional elaboration of items presented on previous trials. This could also explain why in Naveh-Benjamin and Jonides (1984) the RT costs persisted only in the elaboration group but not in the articulatory rehearsal group. We further examined this issue in Experiment 3.

The observation of central attentional demands of articulatory rehearsal clashes with the conclusions reached in a recent study by Vergauwe et al. (2014). Vergauwe and colleagues presented a varying number of verbal memoranda followed by a 12 s processing period. These authors did not observe a delay of responding to the very first CRT in a trial for memory set sizes of up to four words, compared to a 0-words control condition. They assumed that the absence of the set-size effect on CRT was due to participants maintaining the words through articulatory rehearsal. However, in their study, participants were not explicitly instructed to rehearse the words, and no record of their articulatory rehearsal was made. This makes it impossible to verify whether participants in their experiments were actually engaging in articulatory rehearsal, or any other form of rehearsal. In the present experiment, in contrast, participants were explicitly instructed to rehearse the words, and we did observe a rehearsal cost on processing RTs, especially the first one. Our findings suggest therefore that participants in Vergauwe et al.'s (2014) experiments may not have rehearsed the words at all.

One may wonder, however, whether the attentional costs of articulatory rehearsal as estimated in Experiment 1 are specific to the design employed in this study. Experiment 1 did not require immediate recall of the words in the end of the trial, the words were presented simultaneously, and the processing stimuli were presented only sporadically. In addition, participants had to rehearse two words at most, which can be considered as a rather small memory load. This contrasts with the more typical WM set-up used by Vergauwe et al. (2014), which consisted of a Brown-Peterson WM task with a processing task that had to be carried out continuously during the retention interval. To firmly establish the generality of the attentional costs of articulatory rehearsal, in Experiment 2 we assessed the attentional costs of articulatory rehearsal using the experimental set-up of Vergauwe et al. (2014).

Experiment 2

Experiment 1 showed that articulatory rehearsal is not a cost-free strategy. Given the importance of this cost-free assumption in the TBRS theory, in Experiment 2 we sought to replicate the finding of sustained attentional demands of articulatory rehearsal in a more typical WM task set-up. For that purpose, we combined the overt rehearsal protocol of Experiment 1 with a Brown-Peterson WM task as used by Vergauwe et al. (2014). We chose to model Experiment 2 closely on the one of Vergauwe et al. (2014) because the main conclusion from Experiment 1 contradicts their conclusion that articulatory rehearsal does not require central attention.

The goals of the present experiment were twofold. First, we aimed at replicating the attentional costs of articulatory rehearsal in a more typical WM set-up. Second, we increased the maximal number of words to be rehearsed and manipulated set size in a more fine-grained manner by asking participants to hold between 0 and 4 items in WM. This allowed us to better estimate the attentional cost of adding each additional word to the rehearsal set.

Methods

General design

On every trial, a variable number of words had to be retained in WM for a subsequent forward serial order recall test. Thus, different from Experiment 1, participants knew that they had to remember the words over the short term. Set size (from 0 to 4) and the number of syllables of the words in each memory list (mono- or disyllabic) were independently varied as within-subjects factors. Presentation of the words was followed by a 10 s processing phase in which participants responded to a CRT task. Participants were instructed to cumulatively rehearse the memoranda aloud throughout the processing phase. In contrast to Experiment 1, CRT stimuli were presented continuously during the processing phase, imposing a demand on central attention with high temporal density.

Participants

Twenty-one university students (17 women; $M = 24$ years old, $SD = 3.9$ years) took part in Experiment 2. One participant was removed from the final analyses because rehearsals were not recorded due to experimenter error and hence we could not check for compliance with the instructions. This resulted in a final sample of 20 participants.

Materials

For each combination of set size (1–4) and number of syllables (one- or disyllabic), twelve wordlists were constructed, yielding 96 trials. Wordlists were constructed for every participant by randomly selecting words (without replacement) from either a pool of 120 monosyllabic or a pool of 120 disyllabic German words. We added 24 trials with memory load of 0, resulting in a total of 120 trials.

The processing task in Experiment 2 comprised a visuospatial fit judgment task similar to the one employed by Vergauwe et al. (2014). In this task, participants had to judge whether a horizontal bar fitted in the space between two dots (see Fig. 1b). The horizontal distance between the two dots was randomly sampled in every processing episode with replacement from a set of six values. These six values were obtained by dividing the width of the screen by six values equally spaced from 15 to 30 (including 15 and 30). The bar was on each side either 10 pixels longer or shorter than the distance between the two dots. For every decision, it was randomly selected whether the bar appeared above or below the two dots to reduce the probability of the same constellation to appear on several subsequent processing decisions.

Procedure

The sequence of events in each trial is illustrated in Fig. 1b. Every trial began with the presentation of a message announcing the number of words to be remembered on the forthcoming trial. Participants self-initiated the trial by pressing the spacebar. After a 1000 ms blank interval a fixation cross was presented in the middle of the screen for 500 ms, followed immediately by the onset of the first word (for trials with set size > 0). Words were presented sequentially in the center of the screen for 1000 ms. There was no blank interval between the presentations of two successive words. We instructed participants to articulate the words aloud at the time of their presentation to assure that the processing RTs do not reflect the demands of transforming a visual input into a speech plan. After presentation of the last word, they were instructed to rehearse all words aloud in cumulative forward order until the end of the processing period. They were asked to remain silent in trials with 0 words.

Directly after presentation of the last word (or after the fixation cross in 0-word trials), the first processing stimulus was shown. Participants were instructed that their main task was to remember the words in their serial order with high accuracy. At the same time, they should try to respond as fast and accurately as possible to the processing task. The processing period lasted exactly 10 s. Each processing stimulus remained onscreen until participants responded by pressing the left or right arrow keys to indicate a fit or not-fit response, respectively. Each response was followed immediately by the presentation of the next processing stimulus. Participants responded to as many processing stimuli as possible within the 10-s period.

At the end of the 10-s processing period, participants were prompted to recall the words in the order of their presentation using the keyboard. We required the participants to type only the first three letters of each word to minimize their need for typing. They were consecutively cued with the position of the next word to be recalled. The typed letters were shown on the screen, and typos could be corrected by using the backspace key. When satisfied with their answer, participants pushed the return key to confirm their response. Upper and lower case was irrelevant for scoring the responses. The experiment started with 5 practice trials with set size ranging from 0 to 2, which were excluded from the final analyses.

Results

Only RTs in trials in which all words were recalled in their correct serial position at the end of the trial entered the analysis. This criterion resulted in the exclusion of 4.54% of all trials (representing 5.68% of trials with set size > 0). From this pool, RTs associated with incorrect processing responses were excluded (3.44% of remaining RTs). Hence, accuracy in the processing task was generally high. An analysis not reported here showed that accuracy in the processing task did not vary credibly with set size. Next, very long RTs (> 5 s) were also removed (0.01% of the RTs). Mean RTs are plotted against processing position in Fig. 3.

Three sets of responses were used for assessing the costs of articulatory rehearsal over the processing period: (a) responses to the very first processing stimulus in each trial (henceforth referred to as first RTs), which were analyzed separately because they partially reflect the switch costs between encoding the memoranda and preparing for the processing task; (b) average RTs to all subsequent processing stimuli in the trial (henceforth subsequent RTs), which allow for the estimation of the sustained costs of rehearsal if there is one; and (c) response to the very last processing stimulus in each trial (henceforth referred to as last RTs). We analyzed the last RTs separately to test whether central attentional costs of articulatory rehearsal persist even when a substantial amount of rehearsal has happened. Fig. 4a–c shows first, subsequent, and last RTs, respectively, as a function of set size.

Number of syllables was excluded from the analyses, because an initial analysis (not reported here) revealed substantial evidence against an effect of number of syllables (BF = 7.7, 6.7, and 9.1 in favor of the Null for the analysis of first, subsequent, and last RTs, respectively). This initial analysis already implies that any potential costs of articulatory rehearsal are not related to the duration or complexity of articulation (Baddeley, Thomson, & Buchanan, 1975; Service, 1998). We will come back to this issue in the Discussion.

The regression models were run via JAGS. The graphical representation of the models and the respective priors are shown in Fig. 5.

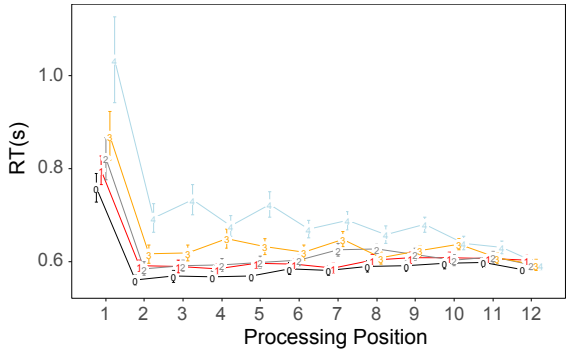


Fig. 3. Mean RTs over processing position presented separately for each set size. Processing position counts the answers to the different stimuli presented during the 10-s processing phase. We only plotted data until processing position 12 because after that there were only few observations. Error bars represent standard errors for within-subjects designs (Bakeman & Mcarthur, 1996).

The model included a fixed effect of set size as a continuous predictor, a random intercept, and a random slope for the effect of set size, allowing the set-size effect to differ between participants (for the importance of including random slopes see Thalmann, Niklaus, & Oberauer, 2017). We approximated the BFs of the fixed set-size effect via the Savage-Dickey density ratio. For models including random slopes, we additionally report the deviance information criterion (DIC, Spiegelhalter, Best, Carlin, & van der Linde, 2002). Here, we report DICs for models including the set-size effect in comparison to models omitting it (see Table 2). The model comparisons support a linear slowing of processing RTs as a function of set size in all three processing positions. Hence, there was clear evidence for a linear set-size effect on processing RTs for the whole 10-s processing period.

Inspection of panels a – c of Fig. 4 suggests that the delaying effect on processing RTs is disproportionately driven by set size 4. Up to set size 3 the set-size effect is much flatter. To compare the present results with the contrast of set-size 0 vs. 2 in Experiment 1, we additionally analyzed whether there is a credible difference between set sizes 0 and 2 for first and subsequent RTs, using the same models as above but entering set size as a categorical predictor. The BFs in favor of the set-size effect were 0.85 and 3.90 for first and subsequent RTs, respectively.[2] The average costs for rehearsing two items were 45 ms and 22 ms for first and subsequent RTs, respectively (i.e., on average 23 ms and 11 ms per additionally rehearsed item). Hence, the additional analyses show that the attentional demand of articulatory rehearsal increases disproportionally when three or more words have to be rehearsed.

A final analysis focused on the question whether articulatory rehearsal is gradually automatized, as suggested by Naveh-Benjamin and Jonides (1984). This gradual automatization would be reflected in a decrease of the set-size effect over processing positions. The second row of Fig. 4 suggests such a decrease: The posterior of the set-size effect on RTs had a smaller mean for subsequent RTs compared to first RTs. However, this tendency most likely appeared because there was large uncertainty about the set-size effect of first RTs. Assessment of the interval covering 95% of the posterior density (hereafter 95% highest density interval, HDI) is informative regarding parameter uncertainty (Kruschke, 2011). It can be seen in the lower panels of Fig. 4 that the HDI was much larger for first RTs compared to subsequent RTs and last RTs. To formally test for a decline of the set-size effect over processing position, we ran a further Bayesian linear model on first, subsequent, and last RTs together, adding processing position (first, subsequent, or last) as further continuous predictor with a random slope. The BF for the main effect of processing position was 14.4, providing strong evidence for slower RTs in the first position compared to subsequent positions. Most importantly, the BF for the interaction of set size x processing position was 0.33, indicating that the Null hypothesis should be favored by a factor of 3. The slight evidence against the interaction between set size and processing position suggests that the attentional costs of articulatory rehearsal rather do not decline over time.

Discussion

Experiment 2 demonstrated in a Brown-Peterson task that articulatory rehearsal delays concurrent processing. This confirms that articulatory rehearsal demands central attention. The analyses showed that (a) the attentional costs are largely due to an uptick in the attentional demand at set sizes 3 and especially 4 and (b) that the attentional costs persisted until the end of the processing period.

The result that up to two words could be rehearsed with negligible costs could be explained by assuming a phonological loop that is limited to the rehearsing of speech of about 2 s length (Baddeley, 2001). Once

[2] The BFs in favor of the set-size effect were 1.85 and 383 for first and subsequent RTs, respectively, when comparing set sizes 0 and 3.

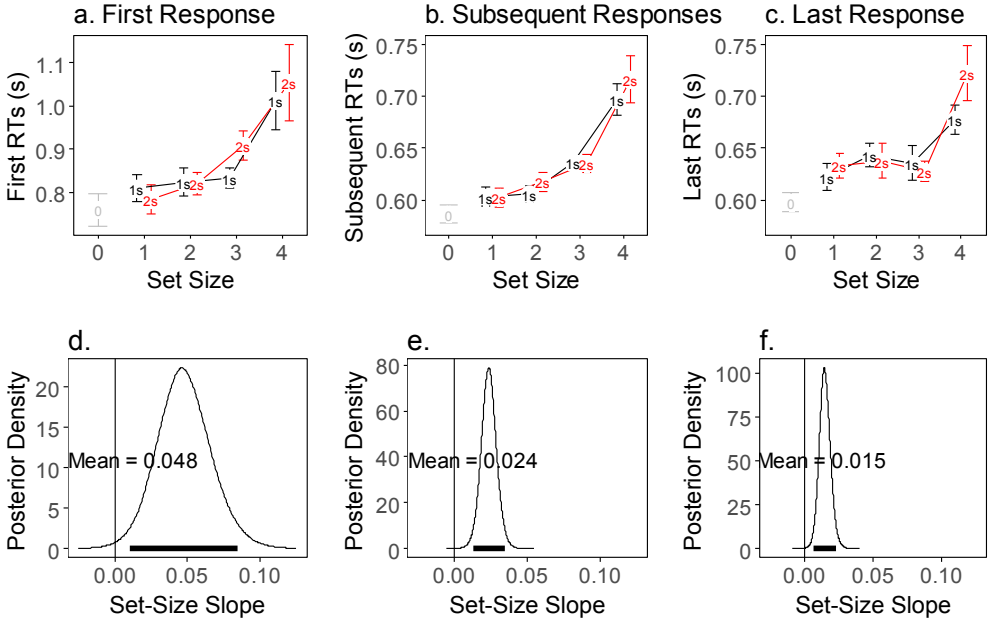


Fig. 4. First RTs (panel a), subsequent RTs (panel b) and last RTs (panel c) for mono- (1 s in figures a. – c.) and disyllabic (2 s in figure a. – c.) words plotted as a function of set size in Experiment 2. Panels d-f show the posterior distributions of the linear set-size effect for first, subsequent, and last RTs, respectively. Error bars in a – c represent standard errors for within-subjects designs (Bakeman & Mcarthur, 1996). The black bars in the bottom of d – f represent 95% HDIs.

the limit is exceeded, an additional, attentionally demanding process (e.g., refreshing of the memory items) is recruited, as proposed by Vergauwe et al. (2014), which would explain the substantial increase in processing RTs from set size 3 on. However, the current data cast doubt on that explanation. A core assumption of the phonological loop model is that long words take longer to rehearse, so that fewer of them can be held in the loop (the word-length effect; Baddeley, 2012; Baddeley et al., 1975). This would have predicted an effect of the number of syllables on processing RTs. Yet, the BF provided evidence against an effect of number of syllables. Therefore, we have to consider different explanations for the pronounced increase of processing RTs with set size 4. One tentative explanation is that the attention participants require monitoring and potentially preventing rehearsal errors increases exponentially with set size. Similarly, rehearsal of a word may also require more attention the less accessible the representation of that word is in WM. That is, rehearsing a word requires retrieving a representation from WM (Tan & Ward, 2000; Watkins & Peynircioğlu, 1982), which is arguably more difficult the more words are stored in WM.

Table 2
BFs and DICs for the models including the set-size effect over the Null model (i.e., omitting it).

	Measure	
Processing position	BF	ΔDIC
First	3	−112
Subsequent	2330	−247
Last	53	−35

Note. ΔDIC (Difference) = DIC (set size) − DIC (Null). Because smaller DIC values reflect better fit, ΔDIC values below zero favor the set-size model.

At first glance, the persistent attentional costs over the whole processing period, and the ambiguous evidence for a set-size effect when only considering set sizes 0 and 2, deviate from the results of Experiment 1. However, the comparison of processing RTs between Experiment 1 and Experiment 2 is not straightforward because the

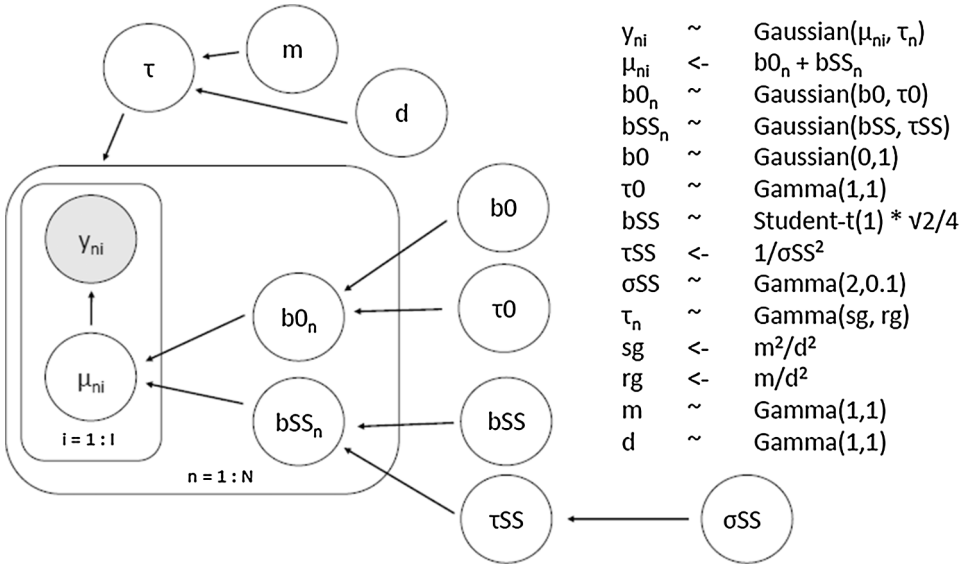


Fig. 5. Graphical representation of the regression models run in JAGS used to predict RTs in Experiment 2. The circle shaded in gray represents the data, the circles without shading represent variables to be estimated. The n-plate indicates independent repetitions over N participants, and the i-plate over I trials. SS = set size.

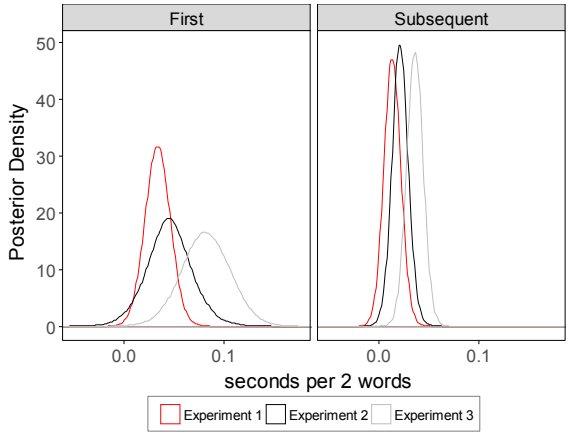


Fig. 6. Posteriors for the difference between rehearsing 0 and 2 words from the data of Experiments 1, 2, and 3 (only from the articulatory rehearsal condition).

presentation of the CRT stimuli was probabilistic in Experiment 1. The first CRT stimulus could appear in any of the five segments. For a comparison between the two experiments we re-analyzed the data from Experiment 1, separating first and subsequent processing responses in each trial, regardless of which segment of the trial the first response occurred in. It can be seen in Fig. 6 that the posteriors of the set-size effects for 0 vs. 2 words largely overlap between experiments. This shows that the results from the two experiments are in agreement with each other, and it suggests that if there are any attentional costs of rehearsing two words, they must be fairly small, particularly during subsequent RTs.

To conclude, Experiment 2 showed that articulatory rehearsal cannot be carried out without any central attentional costs. The costs persisted over a 10 s processing period and they started to magnify especially from set sizes 3 and 4.

Experiment 3

The goal of Experiment 3 was to assess again the attentional costs of rehearsal and elaboration, and in addition to assess whether participants engaged in refreshing, or in elaboration, when rehearsal was prevented through articulatory suppression (AS).

Vergauwe et al. (2014) investigated the set-size effect on processing RTs when people were asked to engage in AS. They assumed that AS forced participants to abandon articulatory rehearsal and to resort to refreshing, a more attentionally demanding maintenance process. They observed substantial set-size effects in the AS condition, which exceeded those in their silent condition. To firmly conclude that AS leads people to use a maintenance strategy more attentionally demanding than articulatory rehearsal, the attentional costs of overt articulatory rehearsal and of maintenance under AS have to be compared in the same experiment. Another possibility is that under AS, participants resort to elaboration rather than to refreshing. Self-report strategy data indicate that participants tend to use elaboration in about 1/3 of the trials in WM tasks (Bailey et al., 2008; Bailey et al., 2011; Dunlosky & Kane, 2007). Elaboration is also assumed to be attentionally demanding. Although Experiment 1 indicated that the attentional costs of elaboration were brief, this strategy may be more demanding when set size is increased, as was the case for articulatory rehearsal in Experiment 2. Experiment 3 was designed to assess the plausibility of these conjectures.

Experiment 3 combined features of Experiments 1 and 2. Participants were assigned to one of two groups (Elaboration or No Elaboration) and their memory for the word lists was assessed again in a surprise delayed-recognition test. There was an immediate memory test in all experimental conditions, and the processing task required responses throughout the processing period.

To investigate the attentional costs of maintenance processes other than articulatory rehearsal, participants were required to perform AS in half of the trials. Hence, participants in the No Elaboration group were instructed to either engage in overt articulatory rehearsal (AR condition), or to perform only AS (AS condition). Participants in the Elaboration group were instructed to either engage in elaboration silently (EL condition), or to perform AS in addition to elaboration (EL + AS condition). Finally, we also varied the concreteness and imageability of the words to be remembered to explore whether elaboration is more beneficial for concrete words with high imageability, which should be easier to elaborate, in comparison to abstract, difficult to imagine words.

This experimental design accomplished three goals. First, it allowed us to replicate the persistent costs of articulatory rehearsal, and its increase with set size. Second, it allowed us to assess whether the costs of elaboration would persist throughout the processing period with increased set sizes, and when a WM test was expected, unlike what we found in Experiment 1. Third, by assessing set-size effects under AS (with or without the instruction to use elaboration), we could measure the putative attentional costs of other maintenance processes participants may resort to in the absence of articulatory rehearsal. Under AS participants may do one of three things: (a) Refreshing, as postulated by Vergauwe et al. (2014). This should lead to a substantial set-size effect on first and subsequent RTs, as observed by Vergauwe and colleagues. (b) Elaboration – this should lead to attentional costs that are similar to the ones observed under the instruction to elaborate under AS. (c) Participants might simply do nothing apart from the requested AS. Overtly articulating irrelevant syllables such as “babibu” could entail its own attentional demand (yielding a main effect of AS), but this cost should not increase with memory load. Hence, in this case we expect no effect of memory set size on CRTs.

The design of Experiment 3 allowed us to test additional predictions arising from possibilities (a) – (c). A summary of the predictions is presented in Table 3. If people are elaborating spontaneously in the AS condition, performance in the AS condition should look similar to the EL + AS condition not only with regard to CRT dual-task costs, but also with regard to delayed recognition and WM performance. Concerning delayed recognition, we expect both elaboration (Craik & Tulving, 1975) and refreshing to increase delayed recognition (e.g., Grillon, Johnson, Krebs, & Huron, 2008; Johnson, Reeder, Raye, & Mitchell, 2002; Johnson et al., 2013; Loaiza, Duperreault, Rhodes, & McCabe, 2014; Loaiza & McCabe, 2012). Therefore, the elaboration conditions should lead to better delayed recognition performance than the AR condition. If people elaborate or engage in refreshing in the AS condition, that condition should also lead to improved delayed recognition, whereas if they do nothing, their delayed recognition should be no

Table 3
Predictions for performance in the AS condition depending on the type of maintenance process used by participants.

	Hypothetical maintenance strategy		
Predicted effect	(a) Refreshing	(b) Elaboration	(c) Nothing
Processing RTs:			
Set-size effect	First RTs and subsequent RTs	= EL + AS	Absent
Memory:			
Delayed test	> AR	> AR = EL + AS	= AR < EL + AS
WM test	?	= EL + AS	< EL + AS
Concreteness effect (immediate and delayed test)	= AR	= EL + AS	= AR

Note. AR = Articulatory rehearsal; EL = Elaboration; AS = Articulatory suppression.

better than in the AR condition.

Concerning WM performance, correlational self-report studies (Bailey et al., 2008; Bailey et al., 2011; Dunlosky & Kane, 2007) suggested that elaboration improves WM performance. Our design allows a first experimental test of that conjecture: WM recall in the EL condition should be better than in the AR condition. If participants in the AS condition engage in elaboration, it should improve WM performance too. At the same time, AS is known to decrease WM performance. Therefore, the beneficial effect of spontaneous elaboration in the AS condition can only be gauged by comparing WM performance in the AS condition to the EL + AS condition: If participants in the AS condition elaborate spontaneously, their WM performance should be comparable to that in the EL + AS condition, because the instruction to elaborate would make little difference to what people do spontaneously during AS. In contrast, if participants in the AS condition do nothing, their WM performance should be worse than in the EL + AS condition. No prediction can be made for the possibility that participants in the AS condition refresh, because we do not know whether refreshing leads to better or worse WM performance than elaboration.

Finally, we aimed to use the concreteness effect as a diagnostic tool to distinguish between different types of rehearsal. If concrete, highly imageable words benefit more from elaboration than abstract, poorly imageable words, participants in the elaboration conditions (EL and EL + AS) should show a larger concreteness effect in immediate and delayed memory tests compared to the AR condition. If this is the case, we can use the magnitude of the concreteness effect to diagnose whether people engage in elaboration in the AS condition.

Methods

Participants

Eighty university students (60 women; $M = 25$ years old, $SD = 4.4$ years) were randomly assigned to one of two groups: No Elaboration ($n = 40$) or Elaboration ($n = 40$). One participant from the No Elaboration group mentioned after the experiment to have expected the delayed memory test. But given that this participant did not mention having used any additional strategy to remember the memoranda, the data were retained for analysis. The data from two subjects (one per group) were excluded. One subject's mother tongue was not German; the other subject already participated in Experiment 2, and the experimenter only realized that after data were collected. These two subjects were replaced by two additional subjects (two women of ages 22 and 23). From the remaining data set four participants were excluded due to low accuracy in the processing task (processing accuracy < 0.65).

Materials

In the present experiment, we manipulated three variables within participants: set size (0, 2, or 4 words), word concreteness and imageability (concrete, highly imageable words vs. abstract, poorly imageable words), and articulatory suppression (without or with AS). For each condition created by the combination of these three variables, eight word lists were generated. Two sets of German words were compiled from the "Semantischer Atlas" data base (Schwibbe, n.d.). One set consisted of words with high ratings of concreteness and imageability (henceforth the concrete word pool, consisting of 96 items), whereas the other set consisted of words with low ratings on both dimensions (abstract word pool, also with 96 items). The word sets were equated for mean word length (mean = 7.8 characters) and frequency (mean log frequency among 4.5 million words = 4.9). For each participant, the memory lists were created by randomly sampling (without replacement) from the respective word pools.

The delayed recognition test was constructed in a similar fashion as described in Experiment 1 with the following exceptions. First, we always selected the first word of a word list as the cue and the second word of that wordlist as the correct alternative to control for serial position across lists with different lengths. There were 64 recognition trials in total – 32 trials for abstract words and 32 for concrete words – that were presented in a randomized order. The correct answer had to be selected amongst an intrusion word from another trial and two new words. Two pools of 64 new words were randomly sampled without replacement from the "Semantischer Atlas" with the only exception that they were not already used for the memory lists. New words for concrete recognition trials were taken from the first pool, new words for abstract recognition trials came from the second pool. The intrusion words selected from another trial could come from any serial position within that trial. Every intrusion probe appeared only once during the recognition test.

Procedure

Experiment 3 combined the procedures of Experiment 2 and Experiment 1. As in Experiment 1, participants were assigned to one of two groups that differed regarding the type of rehearsal instruction (No Elaboration or Elaboration). The sequence of events within a trial was exactly as described in Experiment 2 (see Fig. 1b), with the following exceptions. First, set size was manipulated in a less fine-grained level. Across trials, participants were presented either with 0, 2, or 4 words. Moreover, unbeknownst to the participants, half of the lists consisted of concrete words, and the other half of abstract words. Third, the suppression condition (without or with EL instruction) was manipulated between blocks of trials.

Half of the participants in each group started the experiment with the AS condition (AS or EL + AS), and completed the rehearsal condition without AS (AR or EL) in the second half of the experiment. For the remaining participants, the order of these conditions was reversed. At the beginning of each condition, detailed instructions were displayed on the screen explaining the rehearsal or AS requirements. In the conditions requiring AS, participants were instructed to always articulate "babibu", even when set size was zero. In contrast, in the conditions requiring articulatory rehearsal or elaboration (without AS), participants were instructed to remain silent in 0-words trials. The instruction was followed by three practice trials, one for each memory load (excluded from final analyses).

Results

Delayed recognition

Due to experimenter error the delayed recognition test stopped after 10 trials for four subjects and it stopped after 16 trials for another four subjects. This resulted in only few trials per design cell (set size × word concreteness x AS) for these subjects and therefore we excluded them from the following analysis.

How effective were the different maintenance strategies in laying down an accessible LTM trace? Delayed recognition hit rates are plotted in Fig. 7. They were log-odds transformed before the analysis. The upper part of Table 4 shows the BFs for the contrasts of interest. First, there was a main effect of group, with better delayed memory for the group instructed to elaborate. The beneficial effect of elaboration was also evident when focusing on the two conditions without AS, replicating Experiment 1 and confirming that participants adhered to the elaboration instruction. The latter replicates once more that elaboration is beneficial for long-term memory. Compared to the articulatory rehearsal condition, the beneficial effect of elaboration was also present in the condition where participants elaborated and performed concurrent articulatory suppression (EL + AS). Concrete words were remembered better than abstract words, but the evidence was ambiguous on the question whether elaboration increased the concreteness benefit, in the overall analysis and when only focusing on the two conditions without AS. Comparison of the AS conditions provided no evidence for differences between groups. This indicates that performance in the group told to simply repeat "babibu" aloud (AS condition) and of the group told to repeat this aloud while elaborating (EL + AS) tended to

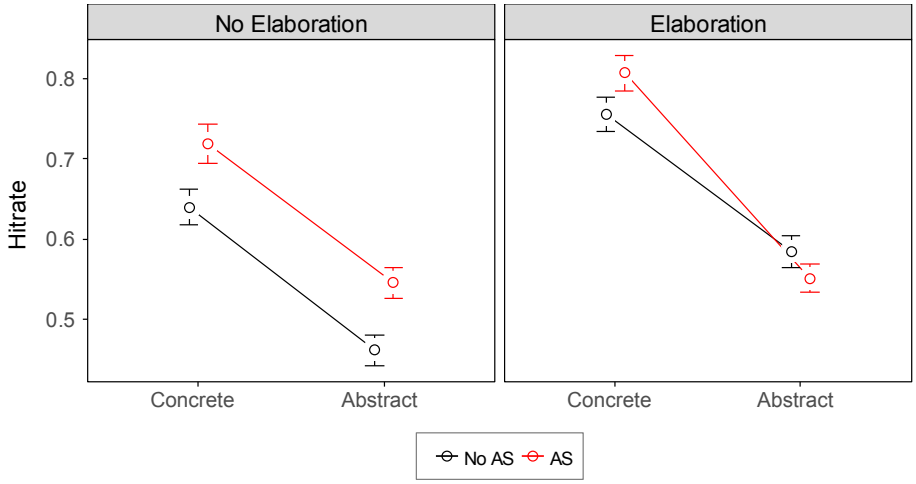


Fig. 7. Average hit rates in the delayed recognition test for concrete and abstract words in the four rehearsal conditions in Experiment 3. Error bars represent standard errors for within-subjects designs (Bakeman & Mcarthur, 1996).

Table 4
Upper panel: Bayes factors for the contrasts of the model predicting proportion of delayed recognition hit rates and several t-tests for pairwise comparisons. Lower panel: Bayes factors for the contrasts of the model predicting serial recall accuracy in the WM recall test (proportion correct) and several t-tests for pairwise comparisons.

Conditions	Effect	BF
Delayed test		
All data	Word concreteness	1.2×10^{20}
	Group (No Elaboration vs. Elaboration)	5.6
	Word concreteness × Group	1.9
No AS conditions	*t*-test: Articulatory rehearsal (AR) vs. Elaboration (EL)	72
	Word concreteness × rehearsal instruction	0.50
AS conditions	*t*-test: Articulatory suppression (AS) vs. EL + AS	0.49
	Word concreteness × rehearsal instruction	2.3
No elaboration	*t*-test: AR vs. AS	18
	Word concreteness × rehearsal instruction	0.43
AR, EL + AS	*t*-test: AR vs. EL + AS	136
WM test		
All data	Word Concreteness	7.3×10^{15}
	AS	6.2×10^{27}
	Group	1.5
	Word concreteness × AS	2.2
	Word concreteness × Group	2.5
	AS × Group	1.9
	Word concreteness × AS × Group	0.95
AS conditions	*t*-test: AS vs. EL + AS	4.8
	Word concreteness × Condition (AS vs. EL + AS)	16

be similar. Fourth, within the No Elaboration group, delayed recognition was worse when participants rehearsed aloud than when they engaged in AS, replicating a result of Camos and Portrat (2015). The finding of better delayed memory in the AS condition suggests that participants used a different maintenance process than articulatory rehearsal in this condition; but it is unclear whether this process is refreshing or elaboration.

WM recall

Our analysis of WM recall was limited to set-size 4, because recall at set-size 2 was very close to ceiling. For brevity, we concentrate on those contrasts that could be informative about which maintenance process participants might have used. Accuracies were log-odds transformed before the analysis (see Fig. 8 for the data and the lower part of Table 4 for the BFs).

Concrete words were remembered better than abstract words, and AS decreased serial recall accuracy. The effect of the elaboration instruction on serial recall accuracy overall was ambiguous. Similarly, the interaction between the elaboration instruction and word concreteness was ambiguous. However, when focusing on the AS conditions, there was strong evidence for the latter interaction, which was driven by a comparatively large concreteness effect in the EL + AS condition. This suggests that participants in the EL + AS condition elaborated the words more than participants in the AS condition.

Processing task

The main dependent variable of interest in this task was RT. Processing accuracy was overall high ($M \sim 95\%$) and there was no evidence for an effect of any of the manipulated variables in this measure (analysis not reported here).

For the analysis of RTs, we only included trials in which participants succeeded in recalling all memoranda in their correct serial position in the WM task. This led to the exclusion of 15.42% of all trials (representing 25.61% of all trials with set size > 0). RTs in the retained trials were further trimmed by removing incorrect responses in the processing task (4.59% of the remaining responses), and by dropping RTs that exceeded 5 s (0.03%). Processing RTs are plotted against processing position for every combination of rehearsal instruction and suppression condition in Fig. 9.

Our main interest was in estimating the attentional costs of each rehearsal strategy as reflected in the slope of the set-size effect. We excluded word concreteness from these analyses because an initial analysis showed that word concreteness had no influence on processing RTs (BFs = 5.72, 5.71, and 5.48 in support of the Null for first, subsequent, and last RTs, respectively). In order to estimate the posterior of the set-size slope in each condition and for each measure of interest (first, subsequent, and last RTs), we entered set size (as a numerical predictor) and rehearsal condition as predictors (including random slopes for these variables) in a hierarchical Bayesian regression model (henceforth called the full model) that was run via JAGS. Specifically, we estimated the set-size slope in each condition and computed the BF for this effect against the Null. In addition, we computed the pairwise comparisons of the set-size slopes between conditions. For all analyses we report the BFs and DICs (see Table 5). The graphical representation of the full model and the respective priors are shown in Fig. 10.

Empirical RT means are plotted as a function of set size in Fig. 11, and the means of the set-size posteriors are shown in Table 6. We observed credible CRT costs in the AR condition that persisted until the end of the processing period, replicating the finding of Experiment 2. The set-size posteriors for first and subsequent RTs only using set sizes 0 and 2 are shown in Fig. 6 for comparison with Experiments 1 and 2. AS

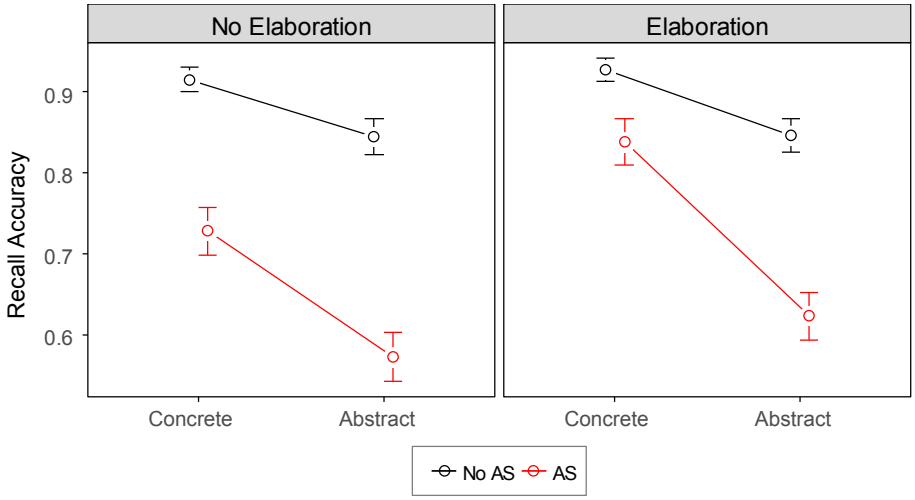


Fig. 8. Mean serial recall accuracy in the WM test as a function of word concreteness in the four rehearsal conditions in Experiment 3. Error bars represent standard errors for within-subjects designs (Bakeman & Mcarthur, 1996).

and EL + AS delayed processing even more than AR in the beginning of processing. However, the costs in these conditions decreased over the processing period and were no longer distinguishable from zero in the last processing episode. The CRT costs in the EL condition were somewhat in between: They were neither clearly distinguishable from those in the AR condition nor from those in the AS and EL + AS conditions. The posterior means of the set-size slopes (Table 6) show that the CRT costs in the EL condition were persistent throughout the

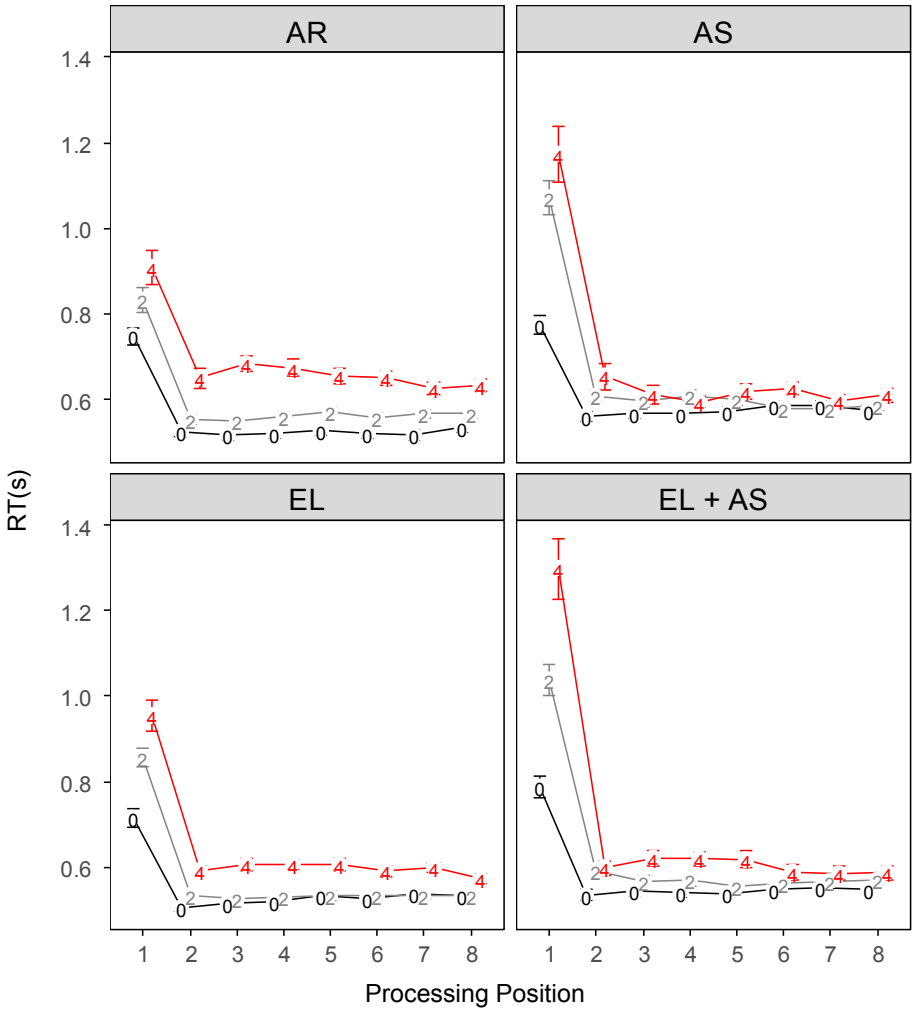


Fig. 9. RTs as a function of processing position. Panels represent the four different rehearsal conditions in Experiment 3. The numbers in the graphs represent the three set-size conditions that are also plotted in different colors. Error bars represent standard errors for within-subjects designs (Bakeman & Mcarthur, 1996).

Table 5
BFs and DICs of the set-size effect in the four conditions (upper table) and of the pairwise comparisons between the set-size slopes in the four conditions.

	Processing Position					
	First		Subsequent		Last	
Comparison	BF	ΔDIC	BF	ΔDIC	BF	ΔDIC
Set-size effect in each condition						
AR	59	−180	4×10^{10}	−412	1.5×10^{7}	−71
AS	1.7×10^{9}	−253	3.8	−91	0.4	1
EL	38,949	−220	2.1×10^{12}	−211	3,496	−40
EL + AS	313,319	−346	24	−93	0.071	3
Pairwise comparison of the set-size effect between conditions						
AR vs. EL	0.09	1	7.1	−9	*0.28*	*−5*
AR vs. AS	60	−71	112	−153	45	−19
AR vs. EL + AS	7.4	−4	33	−3	327	−12
EL vs. AS	7.2	−6	*0.2*	*−3*	*0.57*	*−4*
EL vs. EL + AS	1.6	−118	*0.1*	*−6*	*1.6*	*1*
AS vs. EL + AS	0.11	1	0.07	2	0.051	3

Note. ΔDICs in the upper half of the table represent DIC (set size model) – DIC (Null model). Hence, values below zero favor the set-size model. ΔDICs in the lower half of the table represent the DIC difference between a model that allows the set-size slopes to differ between the two conditions to be compared and a model in which the difference is 0. Hence, values below zero favor the model in which the slopes differ between conditions. Italic values indicate divergence between the BF and the DIC measures.

retention interval. However, compared to the AR condition they were numerically larger in the beginning of processing but numerically smaller in the end of processing.

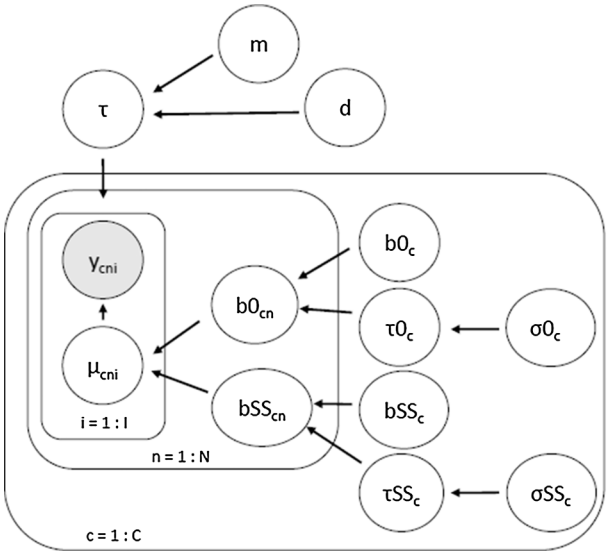
This pattern does not align nicely with the idea that only elaboration has persistent central attentional costs but articulatory rehearsal does not (Naveh-Benjamin & Jonides, 1984). We evaluated the hypothesis of Naveh-Benjamin and Jonides in an additional analysis, focusing only on the AR and EL conditions, and testing whether the CRT cost (i.e., the set-size effect) decreases from first to last RTs more for elaboration than for articulatory rehearsal. RTs were predicted by set size, condition (AR vs. EL), and processing position (first or last RTs) and all their interactions. Random effects were specified for all main effects as well as the intercept. The three-way interaction is shown in Fig. 12 as the two-way interaction between condition and position contrast on the set-size slopes. The BF for the three-way interaction was 7.4, suggesting a tendency that the costs in the EL condition decreased more over time than in the AR condition. This trend is the opposite of the one reported by Naveh-Benjamin and Jonides (1984). Even though this analysis shows a stronger decrease in the EL condition than in the AR condition, it also shows that RT costs in both conditions are persistent over the whole processing period.

Discussion

Experiment 3 shed light on the attentional costs of all three forms of rehearsal: Articulatory rehearsal, elaboration, and refreshing. First, we discuss the findings regarding elaboration and articulatory rehearsal. Then, we consider the most elusive of the three processes, refreshing.

The results of Experiment 3 corroborate the findings of Experiment 1 with regards to the time course of the attentional demand of elaboration and articulatory rehearsal. In both the EL and AR conditions the attentional demand peaked at the first processing RT. Although the set-size slope tended to decrease more in the EL condition than in the AR condition, the attentional costs persisted until the end of the processing period in both conditions. That suggests that setting up elaboration and articulatory rehearsal requires somewhat more central attention than continuously carrying them out. With regards to elaboration it could imply that generating an interactive image of the words is attentionally more demanding than maintaining it. Similarly, setting up a rehearsal scheme for articulatory rehearsal may be attentionally more demanding than carrying out rehearsal itself.

Could the similarities between the EL condition and the AR condition be explained by assuming that subjects in both conditions just used the same maintenance process (i.e., articulatory rehearsal or



y_{cni} ~ Gaussian(μ_{cni}, τ)
μ_{cni} <- $b0_{cn} + bSS_{cn}$
$b0_{cn}$ ~ Gaussian($b0_c$, $\tau0_c$)
bSS_{cn} ~ Gaussian(bSS_c, τSS_c)
$b0_c$ ~ Gaussian(0,1)
$\tau0_c$ <- $1/\sigma0_c^2$
$\sigma0_c$ ~ Gamma(2,0.1)
bSS_c ~ Student-t(1) * √2/4
τSS_c <- $1/\sigma SS_c^2$
σSS_c ~ Gamma(2,0.1)
τ ~ Gamma(sg, rg)
sg <- m^2/d^2
rg <- m/d^2
m ~ Gamma(1,1)
d ~ Gamma(1,1)

Fig. 10. Graphical representation of the regression models run in JAGS used to predict RTs for each condition in Experiment 3. The circle shaded in gray represents the data, the circles without shading represent variables to be estimated. The c-plate indicates independent repetitions over C conditions, the n-plate over N participants, and the i-plate over I trials. SS = set size.

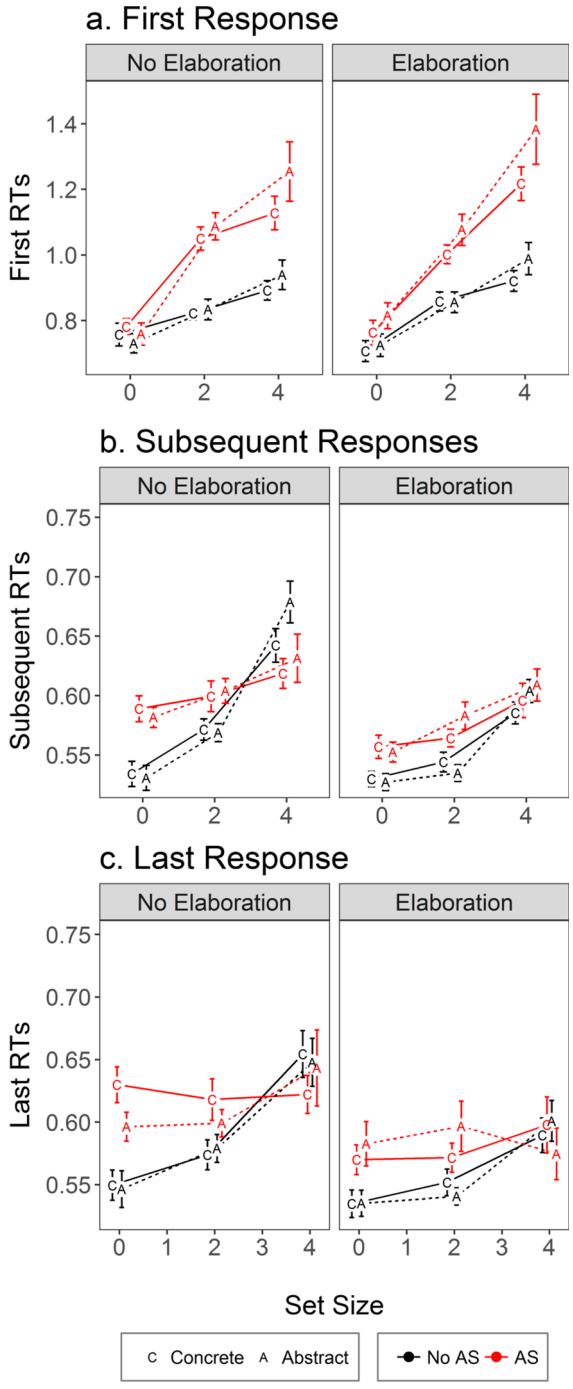


Fig. 11. Empirical RT means as a function of set size for first (panel a), subsequent (panel b), and last (panel c) RTs in Experiment 3. Note the different scales on the y-axes in the panels. Error bars represent standard errors for within-subjects designs (Bakeman & Mcarthur, 1996).

Table 6
Posterior mean slopes of the set-size effect (ms per word) for each of the four rehearsal conditions in Experiment 3.

	Processing position		
	First	Subsequent	Last
AR	44	29	25
AS	122	9	3
EL	58	16	15
EL + AS	117	11	2

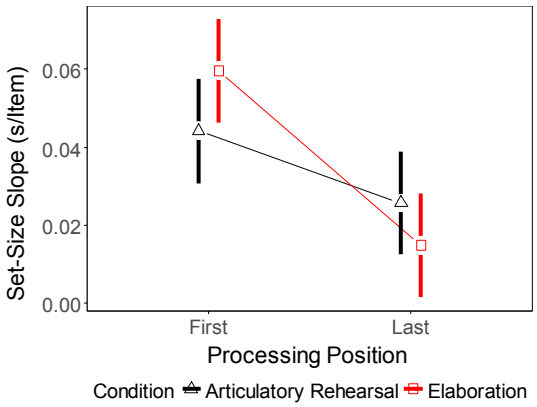


Fig. 12. Posterior means and 95% HDIs of the three-way interaction of rehearsal instruction x position contrast x set size in Experiment 3. The three-way interaction is presented as a two-way interaction of rehearsal instruction x position contrast on set-size slopes.

elaboration)? The results clearly rule this possibility out. As Experiment 1, Experiment 3 showed that delayed recognition was far better in the EL condition than in the AR condition. A simple explanation of that is that subjects in both conditions adhered to the instructions – that is, people in the EL condition elaborated the words, whereas people in the AR condition used articulatory rehearsal.

A tentative explanation for the similar pattern of attentional costs in the EL condition compared to the AR condition is that subjects used articulatory rehearsal in addition to elaboration in the EL condition, hence the parallel. This explanation implies that at least part of the central attentional costs observed in the EL condition is due to articulatory rehearsal. Then, the costs in the EL condition would represent a mixture between those of articulatory rehearsal and those of elaboration. In the extreme case elaboration has no persistent central-attentional costs. This was supported by better delayed memory in the EL + AS condition compared to the AR condition and by the fact that the set-size slope in the EL + AS condition could not be discriminated from zero for last RTs anymore.

What can we learn from Experiment 3 about refreshing? Vergauwe et al. (2014) assumed that AS motivates participants to engage in refreshing of a verbal memory list. This should lead to a substantial effect of memory set size on RTs. In our AS condition we observed a large set-size effect only on the first RTs of each processing period, whereas the set-size effect on subsequent RTs dropped to a low level. This was also observed in the EL + AS condition. Compared to the AR condition, the AS condition was attentionally more demanding only in the initial phase of maintenance. The opposite was true for subsequent and last RTs, in which the demands on central attention were larger in the AR condition than in the AS condition. This contradicts the assumption that participants continuously engage in refreshing during AS throughout the retention interval.

We considered the possibility that participants in the AS condition may use elaboration rather than refreshing. If that was the case, the AS condition should yield comparable effects on CRTs and on memory performance as observed for the EL + AS condition. The evidence for this prediction is mixed. On the positive side we observed similar set-size slopes on CRTs, and both conditions increased delayed recognition relative to the AR condition to about the same degree. On the negative side the concreteness effect in immediate recall was larger with EL + AS than with AS alone. However, the memory results from the WM and the delayed test together do not support the idea that the concreteness effect is meaningfully impacted by elaboration. The current results rather suggest that superior memory for concrete words than for abstract words is due to automatic semantic encoding (Campoy, Castellà, Provencio, Hitch, & Baddeley, 2015; Nittono, Suehiro, & Hori, 2002).

The AS condition tended to result in better long-term memory than the AR condition, a finding replicating a similar observation by Camos and Portrat (2015), and implying that participants in the AS condition did something to improve memory. It appears plausible that participants in the AS condition engage in some attention-demanding process – which may be refreshing or elaboration – briefly at the beginning of the processing interval, which is reflected in their large set-size effect on first RTs, and which caused a modest improvement of long-term memory. In sum, the most parsimonious explanation for our results pertaining to the AS condition is that, after a brief initial effort, participants in this condition engaged in no attention-demanding maintenance process at all.

General discussion

We examined the central attentional costs of articulatory rehearsal and elaboration, and assessed the plausibility that people resort to refreshing when articulatory rehearsal is blocked (i.e., under AS). Prior investigations did not allow conclusions about the central attentional costs of these rehearsal mechanisms for several reasons. The studies of Guttentag (1984) and Naveh-Benjamin and Jonides (1984) used a processing task that did not require response selection. Response selection is critical for measuring central attentional costs (Pashler, 1994). Vergauwe et al. (2014) used a processing task requiring response selection, but participants were not explicitly instructed to perform articulatory rehearsal or refreshing. Therefore, whether and how participants actually rehearsed or refreshed could not be ascertained. We overcame these limitations by using a CRT to measure central attentional costs, and by instructing participants to engage in overt cumulative articulatory rehearsal, or in elaboration.

The primary finding is that articulatory rehearsal clearly requires central attention throughout the retention interval when at least three to four words have to be rehearsed.[3] The fact that rehearsal requires attention confirms the conclusion of Naveh-Benjamin and Jonides (1984). In contrast to these authors, we show that the attentional cost persists for at least 10 s when the memory set consists of more than two or three words.

The fact that about two to three words can be rehearsed with very little costs could be explained by assuming a phonological loop, which has a capacity of about 2 s of speech (Baddeley, 2001). Building on this assumption, Vergauwe et al. (2014) argued that the attentional demands increase from set size 4 because people start to use refreshing once the capacity of the phonological loop is exceeded. Alternatively, it could be that people start to use elaboration in addition to articulatory rehearsal at higher set sizes. Experiment 3 showed that none of these two possibilities is plausible, because conditions in which articulatory rehearsal was blocked did not yield substantial attentional costs. Furthermore, Experiment 2 showed that it is unlikely that the attentional costs of articulatory rehearsal are due to the overload of the phonological loop. A phonological loop account predicts larger attentional costs for disyllabic than for monosyllabic words because longer words take more time to be rehearsed, exceeding the phonological-loop capacity already at smaller set sizes (Baddeley et al., 1975). No such effect was observed in Experiment 2, and Vergauwe et al. (2014; Experiment 5) also did not find an effect of the number of syllables on CRTs. A possibility is that participants increasingly monitor their rehearsal output when set size increases to prevent potential rehearsal errors. Or it could be that rehearsal of more words requires more attention because the accessibility of an individual representation becomes worse the more words have been stored in WM.

The second finding was that whatever people do when instructed to elaborate requires central attention. It is clear that participants instructed to elaborate used elaboration – this is shown by their superior delayed memory compared to articulatory rehearsal. It is less clear that elaboration caused the attentional cost we measured in that group. An alternative interpretation is that people additionally perform articulatory rehearsal when instructed to elaborate, and the persistent attentional demand we observed in the EL condition arose from articulatory rehearsal. This would explain why the costs of elaboration were not persistent in the EL + AS condition, when participants had to elaborate under articulatory suppression that prevented articulatory rehearsal.

Our conclusions regarding the third rehearsal strategy, refreshing, are only based on the results of the AS condition in Experiment 3. Participants in the AS condition appeared to engage in an attentionally demanding strategy briefly after encoding: The set-size slopes on first processing RTs were larger with AS than with articulatory rehearsal. The shallow set-size effect on subsequent RTs implies that memory maintenance in the AS condition demands no central attention. Moreover, the set-size effect completely vanished for the last processing RTs. These results imply that asking participants to engage in AS does not induce a persistent attentionally demanding strategy.

In many regards – in particular the time course of the attentional demand – the AS condition looked similar to a condition with instructed elaboration (EL + AS). Both conditions resulted in better long-term memory than articulatory rehearsal. One noticeable difference between AS with and without elaboration was that instructed elaboration specifically improved WM for concrete words. The same effect was, however, not observed for delayed memory. Therefore, at present we cannot decide whether the spontaneous maintenance strategies participants engage in during AS and the EL + AS are qualitatively different (i.e., refreshing as opposed to elaboration) or merely quantitatively different (namely, less systematic elaboration across trials in the AS condition compared to the EL + AS condition).

Could the slowing of RTs observed in the AR conditions in Experiments 1 and 3 be explained as irrelevant sound effects? Irrelevant sound has detrimental effects on verbal serial recall (Salamé & Baddeley, 1982), so it appears plausible that it could also delay response selection in a CRT task. Three arguments speak against this explanation. First, whereas there is ample evidence for a detrimental effect of irrelevant speech or sound on memory, there is only one study investigating its effect on choice RTs, and that study found no effect (Venetjoki, Kaarlela-Tuomaala, Keskinen, & Hongisto, 2006). Second, distraction by the irrelevant sounds generated through overt rehearsal predicts a main effect of distraction, not the observed effect of memory set size on RT. Third, in the AS condition of Experiment 3 – the only condition in which participants generated speech sound that was actually irrelevant to their task – the set-size slopes vanished quickly over the processing period, in contrast to the persistent set-size slopes in the AR condition.

To conclude, we showed that articulatory rehearsal delays concurrent processing. The RT costs were most noticeable with set size 4 and were persistent until the end of a 10 s processing period. RT costs were initially larger when people were instructed to elaborate, but decreased within a few seconds, and persisted throughout the retention interval only when, in addition to elaboration, participants could also resort to articulatory rehearsal. Finally, preventing articulatory rehearsal through AS does not induce participants to engage continuously in an attention-demanding strategy such as refreshing or elaboration.

These findings require a re-conceptualization of the costs of different rehearsal processes in WM models in three regards. First, because articulatory rehearsal does to some extent require central attention, it cannot be assumed to operate in parallel with refreshing without costs, contrary to what has been proposed by Camos et al. (2009). Rehearsing and refreshing in parallel should be possible when people only have to

[3] Additional evidence that the RT costs can be attributed to central attention from set size 4 on comes from analyses with the EZ diffusion model (Wagenmakers, Van Der Maas, & Grasman, 2007). The analysis scripts together with the data from Experiment 2 and Experiment 3 are available on the OSF page (https://osf.io/69p8j/).

remember two or three words. However, so far no one has claimed that several rehearsal mechanisms are required to remember only two or three words; experiments investigating WM usually involve larger set sizes. Second, contrary to Vergauwe et al. (2014), people do not spontaneously engage in persistent refreshing, or any other central-attention demanding strategy, when maintaining a list of verbal items while engaging in articulatory suppression. Finally, our results indicate that elaboration is initially more demanding than articulatory rehearsal. Whether elaboration has persistent attentional costs is, however, still an open question that we leave open for future research.

Author note

This research was supported by a grant from the Swiss National Science Foundation (project 149193) to K. Oberauer.

References

Baddeley, A. (1996). The fractionation of working memory. *Proceedings of the National Academy of Sciences, 93*(24), 13468–13472.
Baddeley, A. (2012). Working memory: Theories, models, and controversies. *Annual Review of Psychology, 63*(1), 1–29. https://doi.org/10.1146/annurev-psych-120710-100422.
Baddeley, A. D. (1986). *Working memory*. Oxford: Clarendon Press.
Baddeley, A. D. (2001). Is working memory still working? *American Psychologist, 56*(11), 851–864. https://doi.org/10.1037/0003-066X.56.11.851.
Baddeley, A. D., Thomson, N., & Buchanan, M. (1975). Word length and the structure of short-term memory. *Journal of Verbal Learning and Verbal Behavior, 14*(6), 575–589. https://doi.org/10.1016/S0022-5371(75)80045-4.
Bailey, H., Dunlosky, J., & Kane, M. J. (2008). Why does working memory span predict complex cognition? Testing the strategy affordance hypothesis. *Memory & Cognition, 36*(8), 1383–1390. https://doi.org/10.3758/MC.36.8.1383.
Bailey, H., Dunlosky, J., & Kane, M. J. (2011). Contribution of strategy use to performance on complex and simple span tasks. *Memory & Cognition, 39*(3), 447–461. https://doi.org/10.3758/s13421-010-0034-3.
Bakeman, R., & Mcarthur, D. (1996). Picturing repeated measures: Comments on Loftus, Morrison, and others. *Behavior Research Methods, Instruments, & Computers, 28*(4), 584–589. https://doi.org/10.3758/BF03200546.
Barrouillet, P., Bernardin, S., & Camos, V. (2004). Time constraints and resource sharing in adults' working memory spans. *Journal of Experimental Psychology: General, 133*(1), 83–100. https://doi.org/10.1037/0096-3445.133.1.83.
Bartsch, L. M., Singmann, H., & Oberauer, K. (2018). The effects of refreshing and elaboration on working memory performance, and their contributions to long-term memory formation. *Memory & Cognition, 46*(5), 796–808. https://doi.org/10.3758/s13421-018-0805-9.
Brainard, D. H. (1997). The psychophysics toolbox. *Spatial Vision, 10*(4), 433–436. https://doi.org/10.1163/156856897X00357.
Camos, V., Lagner, P., & Barrouillet, P. (2009). Two maintenance mechanisms of verbal information in working memory. *Journal of Memory and Language, 61*(3), 457–469. https://doi.org/10.1016/j.jml.2009.06.002.
Camos, V. (2015). Storing verbal information in working memory. *Current Directions in Psychological Science, 24*(6), 440–445. https://doi.org/10.1177/0963721415606630.
Camos, V., & Barrouillet, P. (2014). Attentional and non-attentional systems in the maintenance of verbal information in working memory: The executive and phonological loops. *Frontiers in Human Neuroscience, 8*. https://doi.org/10.3389/fnhum.2014.00900.
Camos, V., & Portrat, S. (2015). The impact of cognitive load on delayed recall. *Psychonomic Bulletin & Review, 22*(4), 1029–1034. https://doi.org/10.3758/s13423-014-0772-5.
Campoy, G., Castellà, J., Provencio, V., Hitch, G. J., & Baddeley, A. D. (2015). Automatic semantic encoding in verbal short-term memory: Evidence from the concreteness effect. *The Quarterly Journal of Experimental Psychology, 68*(4), 759–778. https://doi.org/10.1080/17470218.2014.966248.
Craik, F. I. M., & Tulving, E. (1975). Depth of processing and the retention of words in episodic memory. *Journal of Experimental Psychology: General, 104*(3), 268–294. https://doi.org/10.1037/0096-3445.104.3.268.
Craik, F., & Tulving, E. (1975). Deth of processing and the retention of words in episodic memory. *Journal of Experimental Psychology: General, 104*(3), 268–294.
Dunlosky, J., & Kane, M. J. (2007). The contributions of strategy use to working memory span: A comparison of strategy assessment methods. *The Quarterly Journal of Experimental Psychology, 60*(9), 1227–1245. https://doi.org/10.1080/17470210600926075.
Greene, R. L. (1987). Effects of maintenance rehearsal on human memory. *Psychological Bulletin, 102*(3), 403–413. https://doi.org/10.1037/0033-2909.102.3.403.
Grillon, M.-L., Johnson, M. K., Krebs, M.-O., & Huron, C. (2008). Comparing effects of perceptual and reflective repetition on subjective experience during later recognition memory. *Consciousness and Cognition, 17*(3), 753–764. https://doi.org/10.1016/j.concog.2007.09.004.
Guttentag, R. E. (1984). The mental effort requirement of cumulative rehearsal: A developmental study. *Journal of Experimental Child Psychology, 37*(1), 92–106. https://doi.org/10.1016/0022-0965(84)90060-2.
Johnson, M. K., Reeder, J. A., Raye, C. L., & Mitchell, K. J. (2002). Second thoughts versus second looks: An age-related deficit in reflectively refreshing just-activated information. *Psychological Science (Sage Publications Inc.), 13*(1), 64.
Johnson, M. R., Higgins, J. A., Norman, K. A., Sederberg, P. B., Smith, T. A., & Johnson, M. K. (2013). Foraging for thought an inhibition-of-return-like effect resulting from directing attention within working memory. *Psychological Science, 24*(7), 1104–1112. https://doi.org/10.1177/0956797612466414.
Johnston, J. C., McCann, R. S., & Remington, R. W. (1995). Chronometric evidence for two types of attention. *Psychological Science, 6*(6), 365–369.
Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association, 90*(430), 773–795. https://doi.org/10.2307/2291091.
Kruschke, J. K. (2011). Doing Bayesian data analysis: A tutorial with R and BUGS.
Lee, M. D., & Wagenmakers, E.-J. (2014). *Bayesian cognitive modeling: A practical course*. Cambridge; New York: Cambridge University Press.
Lewandowsky, S., & Oberauer, K. (2015). Rehearsal in serial recall: An unworkable solution to the nonexistent problem of decay. *Psychological Review, 122*(4), 674–699. https://doi.org/10.1037/a0039684.
Loaiza, V. M., Duperreault, K. A., Rhodes, M. G., & McCabe, D. P. (2014). Long-term semantic representations moderate the effect of attentional refreshing on episodic memory. *Psychonomic Bulletin & Review, 1–7*. https://doi.org/10.3758/s13423-014-0673-7.
Loaiza, V. M., & McCabe, D. P. (2012). Temporal–contextual processing in working memory: Evidence from delayed cued recall and delayed free recall tests. *Memory & Cognition, 40*(2), 191–203. https://doi.org/10.3758/s13421-011-0148-2.
Mora, G., & Camos, V. (2013). Two systems of maintenance in verbal working memory: Evidence from the Word Length Effect. *PLOS ONE, 8*(7), e70026. https://doi.org/10.1371/journal.pone.0070026.
Mora, G., & Camos, V. (2015). Dissociating rehearsal and refreshing in the maintenance of verbal information in 8-year-old children. *Developmental Psychology, 6*, 11. https://doi.org/10.3389/fpsyg.2015.00011.
Morey, R. D., & Rouder, J. N. (2014). Bayes Factor: Computation of Bayes factors for common designs (Version 0.9.7). Retrieved from http://cran.at.r-project.org/web/packages/BayesFactor/index.html.
Naveh-Benjamin, M., & Jonides, J. (1984). Maintenance rehearsal: A two-component analysis. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 10*(3), 369–385. https://doi.org/10.1037/0278-7393.10.3.369.
Nittono, H., Suehiro, M., & Hori, T. (2002). Word imageability and N400 in an incidental memory paradigm. *International Journal of Psychophysiology, 44*(3), 219–229. https://doi.org/10.1016/S0167-8760(02)00002-8.
Pashler, H. (1994). Dual-task interference in simple tasks: Data and theory. *Psychological Bulletin, 116*(2), 220–244. https://doi.org/10.1037/0033-2909.116.2.220.
Pelli, D. G. (1997). The VideoToolbox software for visual psychophysics: Transforming numbers into movies. *Spatial Vision, 10*(4), 437–442. https://doi.org/10.1163/156856897X00366.
Plummer, M. (2003). *{JAGS}: A program for analysis of {Bayesian graphical models using Gibbs sampling. Presented at the Proceedings of the 3rd International Workshop on Distributed Statistical Computing*. Statistical Computing.
Plummer, M., Stukalov, A., & Denwood, M. (2015). rjags: Bayesian graphical models using MCMC (Version 4-4). Retrieved from https://cran.r-project.org/web/packages/rjags/index.html.
Portrat, S., & Lemaire, B. (2014). Is attentional refreshing in working memory sequential? A computational modeling approach. *Cognitive Computation, 1–13*. https://doi.org/10.1007/s12559-014-9294-8.
Development Core, R., & Team (2015). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. ISBN 3-900051-07-0 http://www.R-project.org.
Raye, C. L., Johnson, M. K., Mitchell, K. J., Greene, E. J., & Johnson, M. R. (2007). Refreshing: A minimal executive function. *Cortex, 43*(1), 135–145. https://doi.org/10.1016/S0010-9452(08)70451-9.
Rouder, J. N. (2014). Optional stopping: No problem for Bayesians. *Psychonomic Bulletin & Review, 21*(2), 301–308. https://doi.org/10.3758/s13423-014-0595-4.
Rouder, J. N., Morey, R. D., Speckman, P. L., & Province, J. M. (2012). Default Bayes factors for ANOVA designs. *Journal of Mathematical Psychology, 56*(5), 356–374. https://doi.org/10.1016/j.jmp.2012.08.001.
Rundus, D., & Atkinson, R. C. (1970). Rehearsal processes in free recall: A procedure for direct observation. *Journal of Verbal Learning and Verbal Behavior, 9*(1), 99–105. https://doi.org/10.1016/S0022-5371(70)80015-9.
Salamé, P., & Baddeley, A. (1982). Disruption of short-term memory by unattended speech: Implications for the structure of working memory. *Journal of Verbal Learning and Verbal Behavior, 21*(2), 150–164. https://doi.org/10.1016/S0022-5371(82)90521-7.
Schwibbe, M. H. (n.d.). Der Semantische Atlas. Retrieved from http://kulturkontor-goe.de/semat/semat.htm.
Service, E. (1998). The effect of word length on immediate serial recall depends on phonological complexity, not articulatory duration. *The Quarterly Journal of Experimental Psychology Section A, 51*(2), 283–304. https://doi.org/10.1080/713755759.
Souza, A. S., & Oberauer, K. (2018). Assessing the role of articulatory rehearsal in simple-span and complex-span tasks. [Unpublsihed Manuscript].
Souza, A. S., & Oberauer, K. (2017). The contributions of visual and central attention to visual working memory. *Attention, Perception, & Psychophysics, 79*(7), 1897–1916. https://doi.org/10.3758/s13414-017-1357-y.
Souza, A. S., Rerko, L., & Oberauer, K. (2015). Refreshing memory traces: Thinking of an item improves retrieval from visual working memory. *Annals of the New York Academy of Sciences, 1339*(1), 20–31. https://doi.org/10.1111/nyas.12603.

Souza, A. S., Vergauwe, E., & Oberauer, K. (2018). Where to attend next: Guiding refreshing of visual, spatial, and verbal representations in working memory. *Annals of the New York Academy of Sciences, 1424*(1), 76–90. https://doi.org/10.1111/nyas.13621.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & van der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 64*(4), 583.

Tan, L., & Ward, G. (2000). A recency-based account of the primacy effect in free recall. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 26*(6), 1589–1625. https://doi.org/10.1037/0278-7393.26.6.1589.

Tan, L., & Ward, G. (2008). Rehearsal in immediate serial recall. *Psychonomic Bulletin & Review, 15*(3), 535–542. https://doi.org/10.3758/PBR.15.3.535.

Tombu, M., & Jolicœur, P. (2003). A central capacity sharing model of dual-task performance. *Journal of Experimental Psychology: Human Perception and Performance, 29*(1), 3.

Thalmann, M., Niklaus, M., & Oberauer, K. (2017). Estimating Bayes factors for linear models with random slopes on continuous predictors. *PsyArXiv*. https://doi.org/10.17605/OSF.IO/4XQVR.

Venetjoki, N., Kaarlela-Tuomaala, A., Keskinen, E., & Hongisto, V. (2006). The effect of speech and speech intelligibility on task performance. *Ergonomics, 49*(11), 1068–1091. https://doi.org/10.1080/00140130600679142.

Vergauwe, E., Camos, V., & Barrouillet, P. (2014). The impact of storage on processing: How is information maintained in working memory? *Journal of Experimental Psychology: Learning, Memory, and Cognition*. https://doi.org/10.1037/a0035779.

Vergauwe, E., & Cowan, N. (2014). A common short-term memory retrieval rate may describe many cognitive procedures. *Frontiers in Human Neuroscience, 8*, 126. https://doi.org/10.3389/fnhum.2014.00126.

Vergauwe, E., & Langerock, N. (2017). Attentional refreshing of information in working memory: Increased immediate accessibility of just-refreshed representations. *Journal of Memory and Language, 96*(Supplement C), 23–35. https://doi.org/10.1016/j.jml.2017.05.001.

Wagenmakers, E.-J. (2007). A practical solution to the pervasive problems ofp values. *Psychonomic Bulletin & Review, 14*(5), 779–804. https://doi.org/10.3758/BF03194105.

Wagenmakers, E.-J., Van Der Maas, H. L. J., & Grasman, R. P. P. P. (2007). An EZ-diffusion model for response time and accuracy. *Psychonomic Bulletin & Review, 14*(1), 3–22. https://doi.org/10.3758/BF03194023.

Watkins, M. J., & Peynircioğlu, Z. F. (1982). A perspective on rehearsal. In G. H. Bower (Vol. Ed.), *Psychology of learning and motivation: Vol. 16*, (pp. 153–190). Academic Press. https://doi.org/10.1016/S0079-7421(08)60549-6.

Wetzels, R., Raaijmakers, J. G. W., Jakab, E., & Wagenmakers, E.-J. (2009). How to quantify support for and against the null hypothesis: A flexible WinBUGS implementation of a default Bayesian t test. *Psychonomic Bulletin & Review, 16*(4), 752–760. https://doi.org/10.3758/PBR.16.4.752.