Forgetting is a Feature, not a Bug: Intentionally Forgetting Some Things Helps Us Remember Others by Freeing up Working Memory Resources

Vencislav Popov[1,2], Ivan Marevic[3], Jan Rummel[3] & Lynne M. Reder[1,2]

[1] Department of Psychology, Carnegie Mellon University, Pittsburgh, PA
[2] Center for the Neural Basis of Cognition, Pittsburgh, PA
[3] Department of Psychology, Heidelberg University, Heidelberg, Germany

Correspondence: vencislav.popov@gmail.com

This manuscript has been accepted for publication in Psychological Science. The final published version may differ from the current manuscript

Abstract

We used an item-method directed forgetting paradigm to test whether instructions to forget or remember one item affect memory for subsequently studied items. In two experiments (N_1=138, N_2=33), recall was higher when a word-pair was preceded during study by a to-be-forgotten word-pair. This effect was cumulative: performance increased when more preceding study items were to-be-forgotten. The effect decreased when conditioning memory on instructions for items appearing further back in the study list. Experiment 2 used a dual-task paradigm which suppressed, during encoding, verbal rehearsal or attentional refreshing. Neither task removed the effect, ruling out that rehearsal or attentional borrowing is responsible for the advantage conferred from previous to-be-forgotten items. We propose that memory formation depletes a limited resource that recovers over time, and that to-be-forgotten items consume fewer resources, leaving more available for storing subsequent items. A computational model implementing the theory provided excellent fits to the data.

Keywords: directed forgetting; item-method; directed-forgetting after-effects; computational modeling

I. Introduction

Associative memory formation is an effortful process that can be disrupted by reduced study time (Malmberg & Nelson, 2003), divided attention (Craik, Govoni, Naveh-Benjamin, & Anderson, 1996), or instructions to forget (Bjork, 1972). The probability of forming associative memories decreases with stimulus difficulty – for example, recall and associative recognition are worse for low- compared to high-frequency words (e.g. Criss, Aue, & Smith, 2011; Hulme, Stuart, Brown, & Morin, 2003) and the presence of low-frequency words on a study list hurts memory for other items from the same list (Diana & Reder, 2006; Watkins, LeCompte & Kim, 1998; Malmberg & Murnane, 2002). The ability to form long-term associative memories also depends on working-memory (WM) capacity (Marevic, Arnold, & Rummel, 2018; Unsworth & Spillers, 2010). To explain results like these, we have proposed that binding in memory depletes a limited WM resource that recovers over time (Popov & Reder, 2018; Reder, Liu, Keinath, & Popov, 2016; Reder, Paynter, Diana, Ngiam, & Dickison, 2007; Shen, Popov, Delahay, & Reder, 2018). According to this model, processing weaker items requires more resources than processing stronger items.

Greater demands on limited WM resources means that there are fewer resources available to process additional items. Since the resources recover over time, weaker items within a list especially hurt memory for subsequent items from the same list.

Here, we test a key prediction of the theory – memory should be higher for items that are, during study, preceded by items consuming fewer resources. We used an item-method directed forgetting (DF) paradigm in which each study item is directly followed by either a to-be-forgotten (TBF) or a to-be-remembered (TBR) instruction, indicating whether it will be tested later (Bjork, 1972; Golding & MacLeod, 1998). Previous studies showed worse TBF than TBR recall (i.e., a DF effect), but it is unknown whether memory differs for items that follow a TBR or a TBF item (i.e., a DF after-effect). Investigating the after-effects of memory instructions can shed new light on the role of WM resources for long-term storage.

In line with the Resource Depletion Theory (Popov & Reder, 2018), we propose that, before the remember/forget instructions appear, participants process each item similarly, spending a proportion of their existing resources. After instruction presentation, participants only continue resource-demanding processing of TBR but not TBF items. As a result, fewer resources remain to process items that follow one or more TBR items (compared to one or more TBF items; see Figure S3 in the Supplementary Online Materials, SOM, for an illustration of this prediction).

Early list-method DF research instructing participants to forget a study list before studying a second one supports this idea by showing memory costs for the first but memory benefits for the second list (Bjork, 1970; Epstein, 1972). List-method DF accounts differ regarding the assumed causes for DF costs (e.g., mental context shifts, Sahakyan & Kelley, 2002; Lehman & Malmberg, 2013, or context inhibition, Pastötter, Tempel, & Bäuml, 2017). Most accounts agree, however, that DF benefits are due to participants not rehearsing the preceding TBF list while processing the second list. Yet, different mechanisms might underlie the list-method and item-method DF (Basden, Basden, & Gargano, 1993; Rummel, Marevic, & Kuhlmann, 2016) and it is an open question whether similar beneficial DF after-effects would occur on an item-by-item level. Investigating item-method DF after-effects allows us to further relate the two paradigms and also to characterize this phenomenon with greater detail.

The Resource Depletion Theory makes several predictions concerning DF after-effects. Consider Figure 1 which depicts a study-item sequence. We predict that memory for item $X_{k,}$ $P(X_k)$, will depend on the memory instruction for the preceding items X_{k-i}, where *k* denotes the position of the current item and *i* denotes the lag to the preceding item (e.g. the X_{k-2} item appeared two items ago). Specifically: (1) $P(X_k)$ will be higher when X_{k-1} is TBF rather than TBR; (2) these effects should be cumulative: the more preceding items are TBF, the higher $P(X_k)$ will be; (3) these effects will also depend on the lag *i* between study items: X_{k-1}'s instruction-type effect will be greater than the one for X_{k-2}, etc.

We tested these predictions in two experiments. The first involved a reanalysis of Marevic et al. (2018); the second involved new data from a dual-task experiment which was designed to test whether suppressing rehearsal or dividing attention while concurrently performing the item-method DF task would negate DF after-effects. To show that the Resource Depletion Theory can capture the precise quantitative pattern, we also fit a computational implementation of the account to the data.

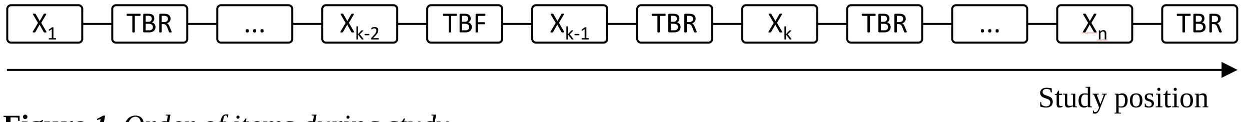

Figure *1*. *Order of items during study*

II. Experiment 1 – Reanalysis of Marevic, Arnold, & Rummel (2017)

A. Method

These methods were described in Marevic et al. (2018) but are also included here to facilitate comprehension of the new information reported herein. The data, materials and analysis code for the current analysis are available at https://github.com/venpopov/directed-forgetting-after-effects.

1. Participants

There were 138 students recruited from Heidelberg University (110 female, M_{age} = 21.96, range: 19-34 years) and they received course credit or monetary compensation. We used the full data set from Marevic et al. (2018), for which the sample-size was originally determined so that it would allow for informative Bayesian decisions regarding the research questions tackled in this article.

2. Materials.

A set of 96 nouns of medium frequency was drawn from the *dlex* database (Heister et al., 2011). Words were randomly paired and assigned to two sets with 24 word-pairs each. One set was used in an initial practice phase and the other was used for the experimental phase. To control for item-specific effects, the assignment of word-pair sets to phases was counter-balanced. In each block, half of the word pairs were followed by TBF and half by TBR instructions. For simplicity, we refer to items followed by TBR instructions as TBR items, and items followed by TBF instructions as TBF items.

3. Procedure.

Experimental sessions started with a working-memory task (not analyzed here but reported in Marevic et al, 2018) and a practice phase in which participants studied 24 TBR and TBF word pairs. Participants were told to only remember the TBR word pairs for a later test and to forget the TBF word pairs. Each word pair was presented for 7 seconds in the center of the screen, followed by either a TBR or TBF instruction for 2 seconds (i.e. the word *remember* or *forget* in German). Trials were separated by a 250-ms inter-stimulus-interval (ISI). After all word pairs had been presented, participants solved math problems for 30 seconds before completing a free recall test. The free recall test was followed by a cued recall test for TBR-items only. Order of recall cues was randomized for each participant. This practice phase was intended to familiarize participants with the paradigm and to increase their belief that the forget instruction was genuine. However, for the real task phase, the procedure was modified so that participants were, again, presented with TBF and TBR items but were asked to recall as many TBR *and* TBF items as possible in the subsequent free and cued-recall tests. Finally, participants performed another working-memory task (not reported), and then were debriefed and received their compensation for participation.

B. Data Analysis

We employed Bayesian statistics for the new analyses of Marevic et al.'s (2018) behavioral data. This approach has several advantages (Wagenmakers, Morey, & Lee, 2016) but most important to us is that Bayes Factors (*BF*s) enabled us to quantify the evidence in favor of the null as well as the alternative hypotheses. We calculated BFs using Bridge Sampling for comparing models that included the effect of interest to models that did not. *BF*s are reported in the direction of the favored model, such that BF_{21} denotes the evidence in favor of model two compared to model one. A *BF* close to 1 means that both models are equally likely, while $BF > 3$ is conventionally interpreted as moderate evidence and a $BF > 10$ as strong evidence in favor of the preferred model (Lee & Wagenmakers, 2013). We applied multilevel logistic Bayesian regressions as implemented in the *brms* R-package (Bürkner, 2017), in which we included crossed random intercepts for subjects and items, as well as random subject slopes for DF effects and after-effects. The population-level regression coefficients had a weakly informative Student *t* distribution prior that was zero-centered with 3 degrees of freedom and scale of 2.5 (Gelman et al, 2008). For the free recall analysis, words were coded as correctly recalled when both items of a pair were recalled. All models were run with 10,000 iterations and 5,000 iterations as burn-in. Convergence was assessed using the potential scale reduction factor $\hat{R}$. For all parameters, $\hat{R} < 1.01$, indicating good convergence.

For each item, we coded whether a TBR or TBF item preceded it. Given that the first item of a study sequence had no predecessor, it was not analyzed. In order to measure the cumulative effect of successive cues, we also coded how many consecutive TBR or TBF items preceded each item. We used a coding scheme that varied from -3 (3 or more consecutive TBF items preceded the current item) to +3 (3 or more consecutive TBR items preceded the current item). For example, if the current study item was preceded by a TBF and a TBR item, in that order, it would have been scored as -1, because there was only one immediately preceding TBF item. Finally, we also looked at the effect of the instructions at each lag individually, without considering other potential intervening items. The output files from the *brms* analyses are available on OSF at https://osf.io/5qd94/files/ under the folder “OSF Storage > analysis_output”.

C. Results

1. Main effect of preceding item type

Figures 2a and *2d* plot the cued and free recall accuracy as a function of the instructions given for the current and the preceding item. There was a DF after-effect, such that both cued and free recall were higher for items that were preceded by TBF items than for those preceded by TBR items (BF_{cued}= 474 and BF_{free}= 3557 for the cued and free recall models with current and preceding instruction type vs. the null model with only current type). There was no interaction between instructions for the preceding item and those for the current item (BF_{cued}= 4.43 and BF_{free}= 17.77 for the cued and free recall models with main effects only against the model with an interaction).

2. Cumulative effect of the number of consecutive preceding TBF or TBR items

Figures 2b and 2e show the cued and free recall accuracy as a function of the number of consecutive preceding TBF or TBR items. Both cued and free-recall performance for the current item were higher when it was preceded by a greater number of consecutive TBF items, and lower, when it was preceded by a greater number of consecutive TBR items**.** The model including the current item’s instructions and the number of consecutive TBF or TBR preceding items fit the data better than the null model that included only the current item’s instructions as a predictor (*BF* = 685 for cued recall and *BF*= 977 for free recall). There was strong evidence that the DF effect and the DF after-effect did not interact (BF_{cued}= 111 and BF_{free}= 100 in favor of the cued and free recall models with main effects only versus the model with an interaction term).

3. Interaction between preceding item type and study position lag

Finally, Figures 2c and 2f plot the cued and free recall accuracy, respectively, as a function of the preceding item type and the lag between that preceding item and the current item on the study list (i.e., ignoring the type for the intervening items). The plots clearly show that the DF after-effect interacted with the lag between the current item and the preceding item – the immediately preceding item had a stronger effect than the one two trials before, which in turn had a stronger effect than the one three trials before. We compared the full model, which included the instructions for items at lags 1, 2, 3 and 4, to identical models without the factor of interest. The posterior parameter estimates from the final model and the corresponding *BF*'s are reported in Table 1 for cued recall and Table 2 for free recall. The DF after-effect from lag 1 was greater than the DF after-effect from lag 2 for both cued and free recall, and the after-effect from lag 3 was greater than the one from lag 4 for cued recall (see Table 1 and Table 2 – parameter comparisons).

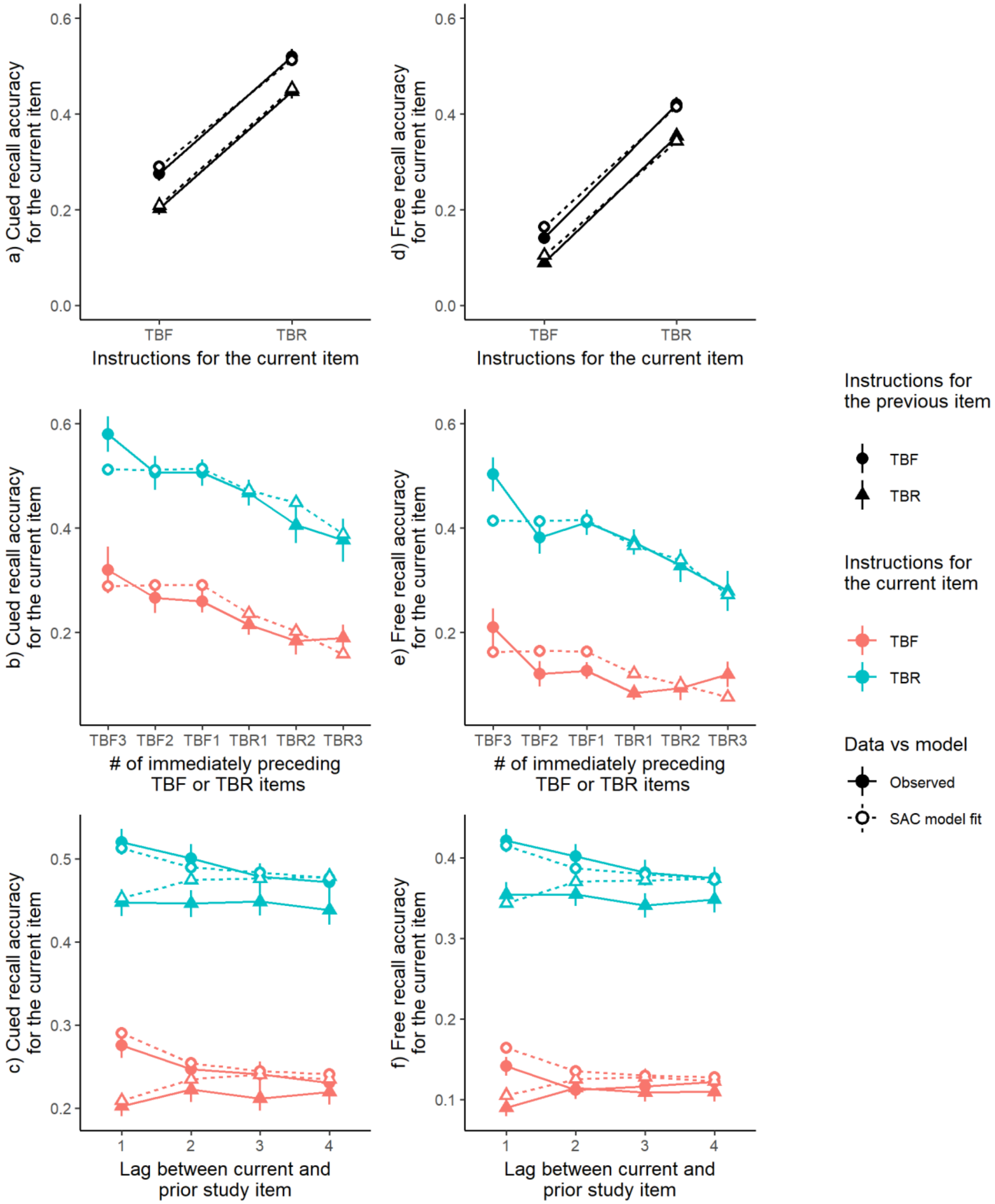

Figure 2. *Results of Marevic et al. (2018) reanalysis and fit of the SAC model – cued recall (a,b,c) and free recall (d,e,f) for the current item, depending on: a), d) whether it was a to-be-remembered (TBR) or to-be-forgotten (TBF) item and whether it was preceded during study by a TBR or a TBF item; b), e) how many of the immediately preceding items during study were TBR or TBF; c), f) what was the study position lag between the current and the prior item (e.g., how many trials ago did the previous item occur). Error bars represent ±1 SE. Solid points and lines represent the data, the empty points and dashed lines represent the predictions of the SAC model.*

Table 1 Parameter estimates for the Bayesian mixed-effects logistic regression of **cued recall**

Fixed-effects	**β**	**Odds Ratio**	**Odds ratio 95% *BCI***	***BF*^**
Intercept (TBF instructions) *	-0.88	0.41	0.30 – 0.58	
TBR instructions for the current item*	1.17	3.21	2.61 – 3.93	4.41×10^{32}
TBR instructions for the item at lag1	-0.41	0.66	0.55 – 0.81	277
TBR instructions for the item at lag2	-0.26	0.77	0.64 – 0.93	3.84
TBR instructions for the item at lag3	-0.23	0.80	0.66 – 0.96	2.61
TBR instructions for the item at lag4	-0.13	0.88	0.73 – 1.05	0.16
Subject random-effects	**σ**	**95% *BCI***		
Intercept	0.79	0.63 – 0.97		
TBR instructions for the current item*	0.50	0.12 – 0.78		
TBR instructions for the item at lag1	0.33	0.02 – 0.68		
Item random-effect	**σ**	**95% *BCI***		
Intercept	0.47	0.32-0.68		
Parameter comparisons	***BF*+**			
Lag1 < Lag2	7.10			
Lag2 < Lag3	1.41			
Lag3 < Lag4	3.63			

Note: Instructions = whether the current item or the items at lag *i* had to be remembered (TBR) or forgotten (TBF). The parameter estimates reflect the means of the posterior distribution. BCI = Bayesian Credible Interval. * indicates models for which the reference category was TBF instruction, so the parameter estimates of the memory instruction effects reflect the odds for correct recall with TBR instructions; ^ Bayes Factor (*BF*) for the model that includes the parameter versus a model that does not. + the Bayes Factor (*BF*) evidence for the difference between the directed forgetting after-effect at different lags.

Table 2 Parameter estimates for the Bayesian mixed-effects logistic regression of **free recall**

Fixed-effects	**β**	**Odds Ratio**	**Odds ratio 95% BCI**	***BF***^
Intercept (TBF instructions) [*]	-1.95	0.14	0.10 – 0.20	
TBR instructions for the current item[*]	1.58	4.88	6.82 – 6.26	3.52×10^{82}
TBR instructions for the item at lag1	-0.49	0.61	0.48 – 0.77	397
TBR instructions for the item at lag2	-0.19	0.83	0.67 – 1.02	0.63
TBR instructions for the item at lag3	-0.22	0.80	0.65 – 0.99	0.78
TBR instructions for the item at lag4	-0.19	0.83	0.67 – 1.02	0.20
Subject random-effects	**σ**	**95% *BCI***		
Intercept	0.30	0.03 – 0.56		
TBR instructions for the current item[*]	0.46	0.10 – 0.73		
TBR instructions for the item at lag1	0.46	0.06 – 0.82		
Item random-effect	**σ**	**95% *BCI***		
Intercept	0.34	0.19 – 0.53		
Parameter comparisons	***BF***+			
Lag1 < Lag2	40.32			
Lag2 < Lag3	0.69			
Lag3 < Lag4	1.45			

Note: Instructions = whether the current item or the items at lag *i* had to be remembered (TBR) or forgotten (TBF). The parameter estimates reflect the means of the posterior distribution. BCI = Bayesian Credible Interval. [*] indicates models for which the reference category was TBF instruction, so the parameter estimates of the memory instruction effects reflect the odds for correct recall with TBR instructions; ^ Bayes Factor (*BF*) for the model that includes the parameter versus a model that does not. + the Bayes Factor (*BF*) evidence for the difference between the directed forgetting after-effect at different lags.

4. SAC computational model of results.

Figure 2 also shows the fit of the SAC Resource Depletion Model. A full description of the model is available in the SOM and in Popov & Reder (2018); we will describe it only briefly and note which of the model assumptions were specifically adapted for this study.

Our model posits that semantic, episodic and contextual information are represented as a network of interconnected nodes that vary in strength. Each node has a current activation value that increases when a node is perceived or when it receives activation from other nodes. This activation decays with time according to an exponential law to a base-level strength of the node. The base-level strength also increases with experience and decreases with time according to a power law. When new information is studied, two processes occur. First, the current and the base level activation values of the preexisting concept nodes are increased. Second, if this is the first occurrence of the study event, a new event node is created, and it gets associated with the corresponding concept and context nodes. If, however, the study event has occurred previously, the existing event node and its links associated with the concept and context nodes are strengthened instead.

During cued-recall, the activation of the list context node and the cue word concept node are raised, which then spread activation to all nodes to which they are connected. The amount of activation that is spread from a node to any given association is multiplied by the strength of its association and divided by the sum total strength of all associated links that emanate from that node. If the current activation of an event node that is connected to the cue concept node surpasses a retrieval threshold, then the correct target word is recalled. The model was not designed to model free recall; however, we simulate free recall by providing only the context node as a cue and evaluating the activation level of all items simultaneously. We also assume that there is output interference during free recall, which we simulate by exponentiating the activation values – this results in squashing the activation of weak items compared to stronger items.

The model also includes a resource pool that is used every time a node is retrieved, created or strengthened. The resource cost of strengthening a node is equal to the degree to which a node is strengthened. Similarly, the resource cost of retrieving a node is equal to the amount of activation necessary to reach the retrieval threshold. During study, if the currently available resource pool is sufficient for storing an item, the memory trace is built or strengthened by the default learning rate. However, if there are currently fewer resources available than required, the memory trace is strengthened proportionally to the remaining resources. The resource pool recovers at a linear rate until it reaches the maximum WM resource capacity.

For the current experiment, we assumed that when an item appears, an episode node is created with a default base-level strength, regardless of the instruction type. Then, when the instruction appears, the episode node for TBR items is strengthened again, whereas the node for TBF items is not. We fit the model by simulating data for each subject, given their specific trial sequence. Six parameters were optimized by minimizing the root mean squared error of the cued recall and free recall data averaged over all subjects, the current instruction type and the number of consecutive preceding TBR or TBF items (24 data points; Figure 2b/e). In our initial modeling, we estimated separate learning rates for the strengthening during item and instruction presentation. These two estimates were roughly equal and the model did not fit the data significantly better than the simpler model with a single learning rate for the strengthening during both item and instruction presentation. The final model parameters were the learning rate $\delta = 0.553$, which governs how much the base-level strength of nodes is increased with each exposure, the resource recovery rate $w_r = 0.526$, the retrieval thresholds for cued-recall $\theta_{\text{cued}} = 0.219$ and for free-recall $\theta_{\text{free}} = 0.167$, as well as the standard deviation of the activation noise $\sigma_{\text{cued}} = 0.831$ and $\sigma_{\text{free}} = 0.431$. All remaining parameters had the default values we have used in prior models. The model provided very good fits to the cued recall (*RMSE* = 0.026, $R^2 = 0.963$) and free recall data (*RMSE* = 0.034, $R^2 = 0.944$). It is noteworthy that the model also captured the interaction between instruction type and lag (Figure 2c/f), although the parameters were not optimized to fit those data points.

III. Experiment 2

Despite good model fit, there remain alternative explanations for Experiment 1's results. People may rehearse or reactivate the memory traces of preceding items *while processing the current item* (Camos, Lagner, & Barrouillet, 2009; McFarlane & Humphreys, 2012). Such rehearsal or attentional borrowing is more likely when the preceding item was TBR rather than TBF (Bjork, 1970) resulting in diminished processing for the current item. Similarly, the REM model (Gillund & Shiffrin, 1984; Lehmann & Malmberg, 2013) postulates that there is a limited rehearsal buffer and that memory trace strength depends on how much of the buffer is currently available. REM would attribute the DF after-effect to the fact that TBF items are not rehearsed, which frees buffer space for the rehearsal of the current item.

In Experiment 2, we tested whether suppressing rehearsal during study would eliminate the DF after-effect to rule out that it is due to greater rehearsal of preceding TBR items (for a similar argument concerning the effect of articulatory suppression on rehearsal-based explanations for the regular DF effect, see Hourihan, Ozubko & Macleod, 2009). We further tested whether the DF after-effect would be attenuated under dived attention to rule out that it is due to allocating attention to previous pairs instead of the current pair (see Figures S4 and S5 of the SOM for illustrations). A stable DF after-effect under suppressed rehearsal or divided attention would support the resource depletion explanation.

A. Method

The rationale, method and parts of the analyses for this experiment were pre-registered at the Open Science Framework (available at https://osf.io/b45tn/). The analysis has changed from the pre-registration from a Bayesian ANOVA to a Bayesian logistic regression, because ANOVAs are not appropriate for analyzing proportion data. The parametric predictions were not included in the pre-registration report. This makes them exploratory for Experiment 1, but confirmatory for Experiment 2. The data, materials and analysis code are available at https://github.com/venpopov/directed-forgetting-after-effects.

1. Participants

Course credit or monetary compensation were given to 33 students from Heidelberg University (22 female, M_{age} = 22.36, range: 18-31 years) who participated in individual sessions. We preregistered this experiment with sample-size requirements of at least 16 participants based on a-priori considerations of statistical power. In order to have enough observations for computational modeling approaches we nevertheless decided to collect more data before we ever looked at the data. As our initial power considerations were based on the assumption that we would conduct a 2 × 4 ANOVA they are also not compatible with the Bayesian logistic regression we used for the final analysis. However, all Bayes factors we calculated provided clear evidence in favor of either the alternative or the null hypothesis, implying that the present sample size was large enough to allow for meaningful conclusions from the present data.

2. Materials

Words of medium frequency were selected from the *dlex* database (Heister et al., 2011), 448 in all, so that they could be randomly paired to form 224 word pairs. The task was divided into eight task blocks. Each block consisted of 12 TBF and 12 TBR word pairs. The memory instructions for individual item pairs were randomized for each participant. The first four items (two TBF, two TBR) of each block served as primacy buffers and were not included in the analyses.

3. Procedure

Participants first received general instructions for the DF task asking them to only remember items that were followed by TBR instructions, but to forget those followed by TBF instructions. Participants were informed that they were about to complete eight study-test blocks of this task while performing a different secondary task in each block. At the beginning of each block, the respective secondary task was explained (see below). Then, each block featured a study phase, in which 12 TBF and 12 TBR items were presented sequentially with a random permutation of the item type order. All other aspects of the *main study procedure* were identical to Experiment 1. During study, participants performed different secondary tasks, which changed every two blocks. The order of secondary tasks was systematically varied across participants using a Latin Square (see Table 3).

Table 3 Counterbalancing orders for the four experimental conditions according to a balanced Latin Square Design

	Block 1 & 2	Block 3 & 4	Block 5 & 6	Block 7 & 8
Order 1	Reh	Att	Reh + Att	Control
Order 2	Att	Control	Reh	Reh + Att
Order 3	Control	Reh + Att	Att	Reh
Order 4	Reh + Att	Reh	Control	Att

Note: Each row represents a unique order, ensuring that each secondary task was followed and preceded by each other condition at least once. Secondary tasks of the same type were always grouped in two consecutive blocks. Reh = rehearsal suppression task, Att = divided attention task, Reh + Att = combined rehearsal suppression and divided attention task, Control = control condition with no secondary task.

In the control blocks, no secondary task was added to the study phase. For the rehearsal suppression blocks, participants were continuously presented via headphones with 60-beats-per-minute metronome sounds and were asked to say the German word “der” [the equivalent word to “the” in English] aloud every time they heard the metronome. Additionally, they had to press the j-key (f-key) or f-key (j-key) whenever saying “der,” to keep the motor component equal across blocks. The assignment of keys was counterbalanced across participants. For the divided attention blocks, participants were continuously presented via headphones with even and odd two-digit numbers. They had to press the j-key for even and the f-key for odd numbers (key assignment counterbalanced). A new number was presented every 2000 ms on average but inter-stimulus-intervals varied between 1250 and 2750 ms to avoid habituation. For the combined rehearsal suppression and divided attention task, participants were also presented with even and odd two-digit numbers but made verbal odd/even judgements. Additionally, they had to press the j or f-key (counterbalanced) with each judgment to align motor demands to the other secondary tasks. The experimenter was present during the entire session, and monitored the compliance with the secondary task – if participants stopped performing the secondary task, the experimenter reminded them to continue engaging with it.

This divided attention task was designed to reduce the attention paid to the main task, but without requiring participants to remember the numbers. In contrast to the resource depletion explanation, which proposes that different amount of resources are depleted *at time t-1*, the attention borrowing explanation implies that the effect is *retroactive* – that is, during the current trial at time *t* participants redirect attention back to the item presented at time *t-1.* The divided attention task would remove the DF after effect in the latter, but not in the former case (see the SOM for more information).

Following each block's study phase, participants always solved math problems for 30 seconds before they performed a free recall test. For these tests, they were always asked to recall as many TBR items as possible in two minutes. We did not ask participants to recall TBF items because there were multiple study-test blocks and thus a TBF recall instruction would not have come as a surprise after the first block. Participants were specifically encouraged to recall both words of the pairs if possible, but if they could recall only one word of the pair, they should report it as well. Then, participants performed a cued-recall test for which they were presented with the first words of all TBR item pairs they had studied (in random order) and were asked to recall the second word. After four blocks, participants were given a three-minute break in which they received water but had to stay in the laboratory. After completing all eight blocks, participants were asked whether they used a certain forgetting strategy and some demographic questions.

B. Results

1. Main effect of preceding item type and dual task condition.

Figures 3a and 3d plot the cued and free recall accuracy as a function of the memory instructions for the preceding item and the dual-task condition. Both cued and free recall were higher for items that were preceded by TBF items rather than TBR items (BF_{cued} = 13 and BF_{free} = 134 for the cued and free recall models with dual-task condition and preceding instruction type vs. the null model with only dual-task condition as a factor). Overall, memory performance was lower in all dual-task conditions compared to the control condition (BF_{cued} = 411 and BF_{free} = 500 for the model with dual-task condition as main factor, against the null model). This overall memory decline indicates that the dual task conditions was effective in preventing participants from engaging in rehearsal and/or refreshing during study. Nevertheless, the DF after-effect was present in all conditions, since the preceding items' instructions did not interact with dual-task condition (BF_{cued} = 395 and BF_{free} = 1515 for the models with main effects only against the models with an interaction). Because the main effect of preceding instruction type did not differ between conditions, we report all remaining analyses collapsed over conditions.

2. Cumulative effect of the number of consecutive preceding TBF or TBR items.

Figures 3b and 3e show the cued and free recall accuracies as a function of the number of consecutive preceding TBF or TBR items. Both cued and free recall performances for the current item were higher when it was preceded by a greater number of consecutive TBF items and lower when it was preceded by a greater number of consecutive TBR items**.** The model including the number of consecutive TBF or TBR items fit the data better than the null model (BF_{cued} = 1402 and BF_{free} = 99).

3. Interaction between preceding cue and study position lag.

Finally, the DF after-effect interacted with the study lag between the current item and the preceding item – the immediately preceding item had a stronger effect than the one two trials before, which in turn had a stronger effect than the one that occurred three trials before (*Figure 3c/f*). We compared the full model, which included the instructions for items at lags 1, 2, 3 and 4, to identical models without the factor of interest. The posterior parameter estimates from the final model and the corresponding *BF*'s are reported in Table 4 for cued recall and Table 5 for free recall.

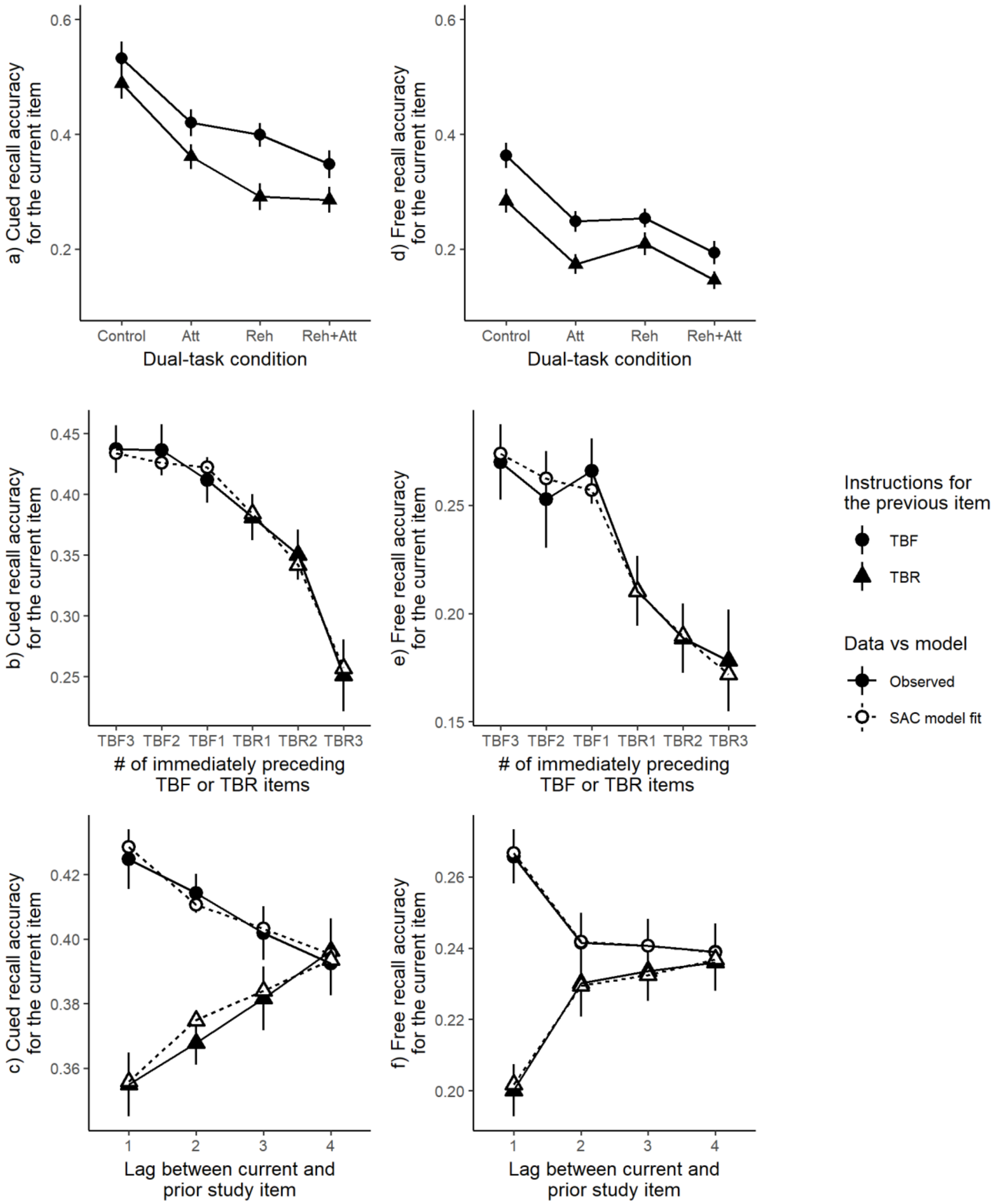

Figure 3 *Results of Experiment 2 and SAC model fits – cued recall (a,b,c) and free recall (d,e,f) for the current item depending on (a, d) whether it was preceded during study by a TBR or a TBF item and the dual task condition (Control = No dual task, Att = Divided attention, Reh = suppressed rehearsal, Reh+Att = simultaneous divided attention and suppressed rehearsal; (b, e) how many of the immediately preceding items during study were TBR or TBF; (c, f) what was the study position lag between the current and the prior item (e.g., how many trials ago did the previous item occur). Error bars represent ±1 SE.*

Table 4 Parameter estimates for the Bayesian mixed-effects logistic regression of **cued recall**

Fixed-effects	**β**	**Odds Ratio**	**Odds ratio 95% *BCI***	***BF*^**
Intercept (TBF instructions; Control) *	0.44	1.56	0.92 – 2.67	
Effects of dual-task condition				
Divided attention (DA) condition	-0.66	0.52	0.31 – 0.87	177.57
Suppressed rehearsal (SR) condition	-0.54	0.43	0.26 – 0.71	1874
DA + SR condition	-1.13	0.32	0.19 – 0.54	> 15 × 10^3
Effects of instructions				
TBR instructions for the item at lag1	-0.39	0.68	0.54 – 0.85	17.94
TBR instructions for the item at lag2	-0.28	0.76	0.62 – 0.92	2.93
TBR instructions for the item at lag3	-0.15	0.86	0.71 – 1.05	0.18
TBR instructions for the item at lag4	-0.01	0.99	0.81 – 1.20	0.05
Subject random-effects	**σ**	**95% *BCI***		
Intercept (control)	1.14	0.85 – 1.52		
Divided attention	0.65	0.19 – 1.12		
Rehearsal suppression	0.56	0.11 – 1.00		
DA + RS	0.69	0.28 – 1.13		
TBR instructions for the item at lag1	0.28	0.02 – 0.69		
Item random-effect	**σ**	**95% *BCI***		
Intercept	0.91	0.76 – 1.08		
Parameter comparisons	***BF*+**			
Lag1 < Lag2	3.37			
Lag2 < Lag3	4.65			
Lag3 < Lag4	5.57			

Note: Instructions = whether the items at lag *i* had to be remembered (TBR) or forgotten (TBF). * the reference category was when the preceding item had forget instructions, so the parameter estimates of the instruction effects reflect the odds for correct recall with remember instructions for preceding items; ^ Bayes Factor (*BF*) for the model that includes the parameter vs a model that does not. + the Bayes Factor (*BF*) evidence for the difference between the cue effect at different lags. *BCI* = Bayesian Credible Interval. The parameter estimates reflect the means of the posterior distribution.

Table 5 Parameter estimates for the Bayesian mixed-effects logistic regression of **free recall**

Fixed-effects	**β**	**Odds Ratio**	**Odds ratio 95% *BCI***	***BF***$^\wedge$
Intercept (TBF instructions; Control) *	-0.65	0.52	0.34 – 0.78	
Effects of dual task condition				
Divided attention (DA) condition	-0.77	0.46	0.31 – 0.69	$> 15 \times 10^3$
Suppressed rehearsal (SR) condition	-0.65	0.52	0.34 – 0.79	651.17
DA + SR condition	-1.15	0.32	0.19 – 0.51	$> 15 \times 10^3$
Effects of instructions				
TBR instructions for the item at lag1	-0.48	0.62	0.47 – 0.81	30.53
TBR instructions for the item at lag2	-0.12	0.89	0.72 – 1.10	0.15
TBR instructions for the item at lag3	-0.09	0.92	0.75 – 1.13	0.05
TBR instructions for the item at lag4	-0.08	0.92	0.74 – 1.14	0.06

Subject random-effects	**σ**	**95% *BCI***
Intercept (control)	0.63	0.42 – 0.90
Divided attention	0.21	0.01 – 0.58
Rehearsal suppression	0.44	0.04 – 0.86
DA + RS	0.67	0.19 – 1.18
TBR instructions for the item at lag1	0.38	0.03 – 0.77

Item random-effect	**σ**	**95% *BCI***
Intercept	0.70	0.54 – 0.87

Parameter comparisons	***BF***$^+$
Lag1 < Lag2	69.42
Lag2 < Lag3	1.37
Lag3 < Lag4	1.04

Note: Instructions = whether the items at lag *i* had to be remembered (TBR) or forgotten (TBF). * the reference category was when the preceding item had TBF instructions, so the parameter estimates of the instruction effects reflect the odds for correct recall with TBR instructions for preceding items; ^ Bayes Factor (BF) for the model that includes the parameter vs a model that does not. + the Bayes Factor (BF) evidence for the difference between the cue effect at different lags. BCI = Bayesian Credible Interval. The parameter estimates reflect the means of the posterior distribution.

4. *SAC computational modeling.*

Similar to Experiment 1, we fit the SAC model by simulating data for each subject, given their specific trial sequence. There is no rehearsal mechanism in the model and, for that reason, we ignored the dual-task conditions and only modeled the effect of the prior cue. The same six parameters were optimized by minimizing the root mean squared error of the cued recall and free recall data averaged over the number of consecutive preceding TBR or TBF items (12 data points; Figure 3b/e). In addition, we had to increase the free recall output interference exponent parameter, to account for the different performance in free and cued recall. The estimated parameters were very similar to those of Experiment 1 – learning rate $\delta = 0.639$, resource recovery rate $w_r = 0.551$, the retrieval thresholds for cued-recall $\theta_{\text{cued}} = 0.279$ and for free-recall $\theta_{\text{free}} = 0.457$, and the standard deviation of the activation noise $\sigma_{\text{cued}} = 0.451$ and $\sigma_{\text{free}} = 0.868$. All remaining parameters had the default values we used in prior models. The model provided excellent fits to the cued recall (*RMSE* = 0.008, $R^2 = 0.991$) and free recall data (*RMSE* = 0.005, $R^2 = 0.984$). It is noteworthy that the model also captured the fact that the DF after-effect decreases with lag (Figure 3c/f), even though the parameters were not optimized to fit those data points.

IV. General Discussion

We demonstrated a novel DF after-effect – when an item is to-be-forgotten rather than to-be-remembered memory for the subsequent item benefits. This effect occurs in both cued and free recall; it is cumulative, such that the more preceding items are TBF the higher the memory benefits; the effect decreases when conditioning memory on instructions for items appearing further back in the study list. The DF after-effect was replicable and remarkably consistent across the two experiments – the cued-recall odds ratios associated with items preceded by TBR items relative to TBF items were 0.66 and 0.67, respectively.

Previous research has also shown improved memory for whole lists when a preceding list was TBF rather than TBR (Bjork, 1970; Epstein, 1972). This is, however, the first study to demonstrate DF after-effects on an item level and to characterize in detail how the precise order of TBR and TBF items affects memory for subsequent items. The present findings indicate similarities between the two DF methods but also provide new theoretical insight, because the item-method allows for a more fine-grained investigation of the DF after-effects. For example, researchers have argued that the list-method DF after-effect is due to less rehearsal borrowing (Bjork, 1970; Sahakyan & Kelly, 2002). This explanation is unlikely to hold for the item-method, because the DF after-effects in our experiments were not attenuated when rehearsal was prevented.

What causes the item-method DF after-effects? We propose that memory formation and storage deplete a limited resource that recovers over time (Reder et al, 2007; Popov & Reder, 2018). Within this framework, TBR items deplete more resources, and they leave fewer resources for processing subsequent items. A computational model implementing the theory provided excellent fits to the cued and free recall data. Although we do not know whether DF after-effects would appear in other tasks (e.g., recognition) or with other materials (e.g., single words), DF is not the only manipulation that leads to after-effects – similar patterns occur when the preceding items are of high- rather than low-frequency, or have been repeated more often in the experiment (Popov & Reder, 2018). These other after-effects occur under a variety of encoding and retrieval conditions, and the general pattern is remarkably similar to the one found for DF here. Item-specific after-effects seem to be a general phenomenon that can be tied together with the current model.

The idea that TBR and TBF items differ in the required processing resources is not new. Fawcett & Taylor (2008; 2012) argued that participants actively withdraw attentional resources from TBF items when being presented with a forget instruction, freeing resources to process prior TBR items. The key difference between this research and ours is that, whereas Fawcett & Taylor measured incidental memory for secondary probes presented shortly after the forget instructions and not relevant to the primary memory task, we measured intentional memory for subsequent study items. Fawcett & Taylor found RTs to post-TBF probes to be slower than to post-TBR probes and recognition memory for post-TBF probes to be worse than for post-TBR probes. Fawcett & Taylor (2012) suggest that these effects are indicators of greater processing in the immediate aftermath of TBF compared to TBR instructions. Our experiments were not designed to measure forget-instruction-induced attention withdrawal and thus our findings do not speak for or against the existence of such a process. However, if such an attention withdrawal process existed it would need to be short-lasting and not overly resource taxing. Otherwise, we would not have observed memory benefits from preceding TBF item but rather the opposite.

Are there alternative explanations for the DF after-effect phenomenon? We discount three possibilities. First, the DF after-effect cannot be due to continued rehearsal of preceding TBR items – articulatory suppression makes verbal rehearsal nearly impossible, and it would have eliminated the effect were it due to rehearsal borrowing. Second, if memory for the current item was worse because participants were directing their attention to the preceding TBR items, then dividing attention should have reduced the DF after-effect proportionally to the overall reduction in memory. This prediction follows if we assume that dividing attention makes it less likely that participants use their remaining attentional resources to process preceding items, but that they would rather focus them mostly on the current item (See Figure S5 in the Supplementary Online Materials). Whereas dividing attention reduced recall, the DF after-effect was not attenuated. It is nevertheless possible to imagine alternative formulations of attentional refreshing that might be consistent with this data. A final alternative is that when an item is forgotten, the surrounding items become more distinct and easier to retrieve (e.g., Brown, Neath, & Chater, 2007; Sederberg, Howard, & Kahana, 2008). This explanation would predict that TBR items should impair memory for both preceding and following study items. We did not find support for this prediction – accuracy for the current item did not differ depending on whether it was followed by TBF or TBR items during study (see SOM for details).

The disparity between effects of preceding and subsequent item types distinguishes the DF-after-effect from general distinctiveness effects, in which distinct items impair memory for *all* surrounding items (Detterman, 1975). The fact that memory for the current item was not affected by whether the subsequent item was TBR or TBF also renders a compartmentalization explanation, as suggested by the REM buffer model of Lehmann & Malmberg (2013) for example, less likely. Their model proposes that the presentation of distinct items cause previously studied items to be dropped from rehearsal and that distinct items are more persistent (Kamp, Lehman, Malmberg, & Donchin, 2016). A direct computational comparison of the REM and SAC model predictions would be necessary to adjudicate between the alternative interpretations and presents a venue for future research.

V. Author Contributions

L. M. Reder developed the theory. V. Popov developed the study concept. Data for Experiment 1 came from Marevic et al. (2018). V. Popov, I. Marevic, and J. Rummel designed Experiment 2. Testing and data collection were performed by I. Marevic. I. Marevic and V. Popov performed the data analyses. V. Popov implemented the SAC model. V. Popov and I. Marevic drafted the initial manuscript. J. Rummel and L. M. Reder provided critical feedback and revisions.

VI. Acknowledgements

We are grateful to Matt So for proofing the final manuscript.

VII. Appendix – Online Supplementary Materials

A. Discounting a distinctiveness explanation – the effect of subsequent item type

Is it possible that DF after-effects can be explained by assuming that an item surrounded by TBF items becomes more distinct and suffers less interference from those surrounding items? Postulating temporal distinctiveness plays an important role in numerous models of episodic memory (e.g., Brown, Neath, & Chater, 2007; Sederberg, Howard, & Kahana, 2008). These theories predict not only that memory for the current item should be better when preceded by a TBF item (i.e. the DF after-effect), but also when the *subsequent* item(s) are also TBF.

To test the adequacy of distinctiveness explanations for our data, we re-ran all analyses conditioning memory for the current item on whether participants had to remember or forget the item(s) that followed it. In Experiment 1, memory for the current item did not differ as a function of whether the subsequent item was TBF ($M_{\text{cued}} = 0.38$, $SD_{\text{cued}} = 0.20$, $M_{\text{free}} = 0.27$, $SD_{\text{free}} = 0.14$) or TBR ($M_{\text{cued}} = 0.35$, $SD_{\text{cued}} = 0.19$, $M_{\text{free}} = 0.25$, $SD_{\text{free}} = 0.13$; $BF_{cued} = 28$, $BF_{free} = 25$ in favor of the model without the subsequent item type as a factor). The full data pattern related to the subsequent item type is shown in Figure S1.

In Experiment 2 there was no clear evidence for the presence or the absence of an effect of the subsequent item type on cued recall accuracy ($M_{\text{TBR}} = 0.37$, $SD_{\text{TBR}} = 0.20$, $M_{\text{TBF}} = 0.40$, $SD_{\text{TBF}} = 0.19$; $BF = 2.17$ in favor of the null model without subsequent item type), and any potential effect was not modulated by the divided attention manipulations ($BF = 610$ in favor of the model without an interaction). Free recall in Experiment 2 was numerically slightly better when the subsequent item was TBF ($M = 0.25$, $SD = 0.11$) rather than TBR ($M = 0.22$, $SD = 0.10$), but there was no clear evidence in favor of this effect ($BF = 2.64$). Furthermore, in free recall, the effect of the subsequent item type was less than half the size of the effect for the preceding item type (3% vs 7% respectively when followed/preceded by one TBR or TBF item; 3% vs 10% when followed/preceded by 3 TBR or TBF items; $BF_{\text{preceding>subsequent}} = 11.3$). Figure S2 shows the full data pattern for effects of the subsequent item types in Experiment 2.

In summary, Experiment 1 provided strong evidence that subsequent item types do not affect memory for the current item, whereas in Experiment 2 there was no clear-cut evidence against the alternative explanation that items surrounded by TBF items become more distinct. The potential distinctiveness effect on free recall in Experiment 2 was numerically smaller than the DF after-effect and, unlike the DF after-effect, it was not statistically reliable. Furthermore, virtually no distinctiveness effect was present in the cued-recall data. One could ask whether distinctiveness models that compress time (e.g. SIMPLE; Brown, Neath, & Chater, 2007) would predict this asymmetry in the effect of preceding and subsequent TBF items. SIMPLE suggests that the mental representation of time is logarithmically compressed, such that items further back in a study sequence are closer to each other in mental time (see Figure S3). This assumption might indeed lead to asymmetric effects of preceding and subsequent TBF items, but this asymmetry would be opposite to the one we found in the current study. A preceding item that is not stored will, due to the logarithmic compression of time, create a smaller temporal gap next to the item of interest than a subsequent item that is not stored (see Figure S3, bottom). Even though in real time the duration of the gap would be the same, in compressed time, the preceding gap would be compressed more, since it is further back from the current moment. As a result, distinctiveness models like SIMPLE that compress time representations logarithmically would predict that subsequent TBF items should have a bigger effect on memory for the current item than preceding TBF items. We found exactly the opposite result. Combined, these results suggest that if distinctiveness plays a role it is a minor one at best, and cannot account for the full DF after-effects.

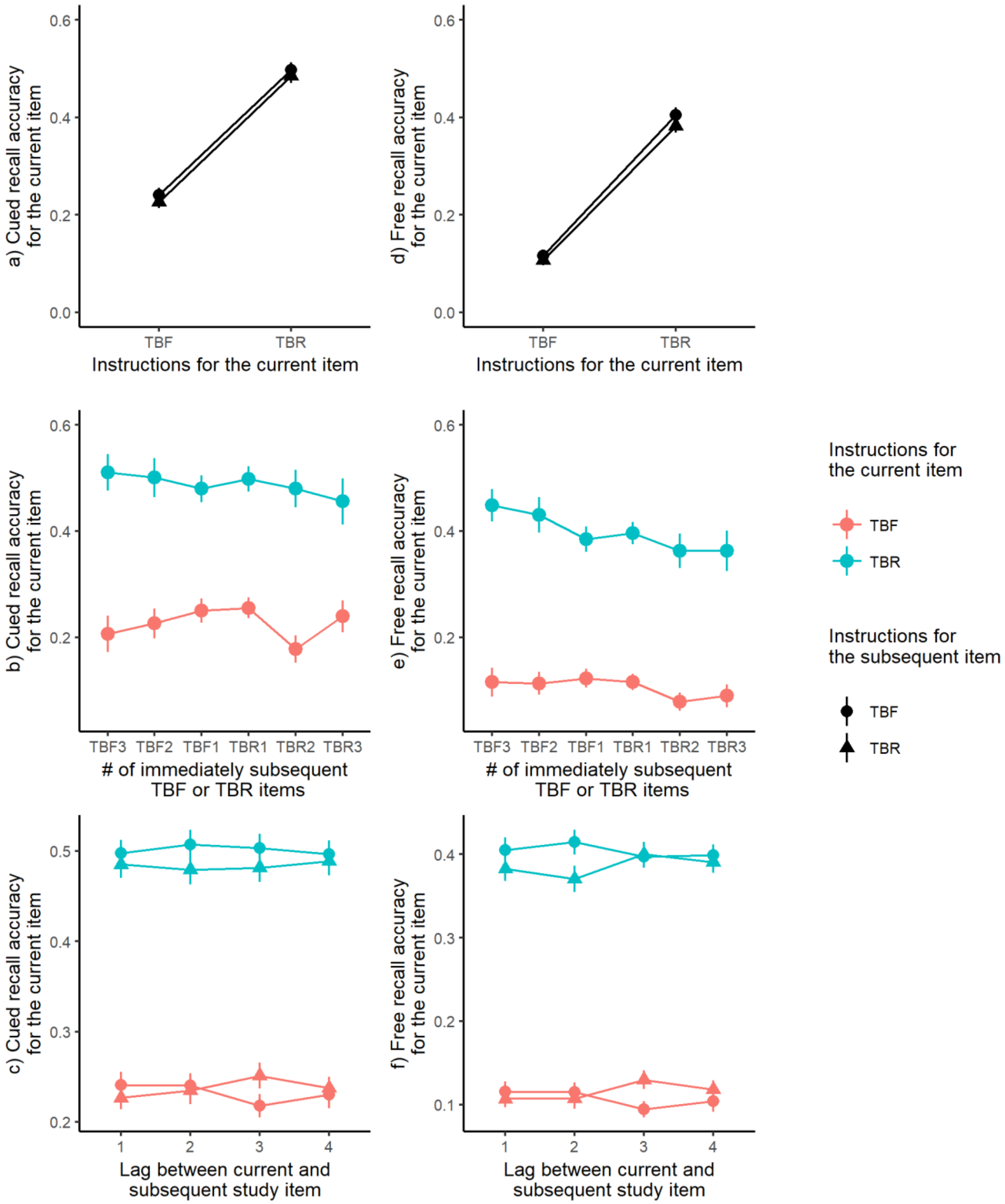

Figure S1. Cued recall (a,b,c) and free recall (d,e,f) for the current item in Exp. 1, depending on: a, d) whether it was a to-be-remembered (TBR) or to-be-forgotten (TBF) item and whether it was ***followed*** *during study by a TBR or a TBF item; b, e) how many of the immediately* ***following*** *items during study were TBR or TBF; c, f) what was the study position lag between the current and the* ***subsequent*** *item. Error bars represent ±1 SE. Solid points and lines represent the data, the empty points and dashed lines represent the predictions of the SAC model.*

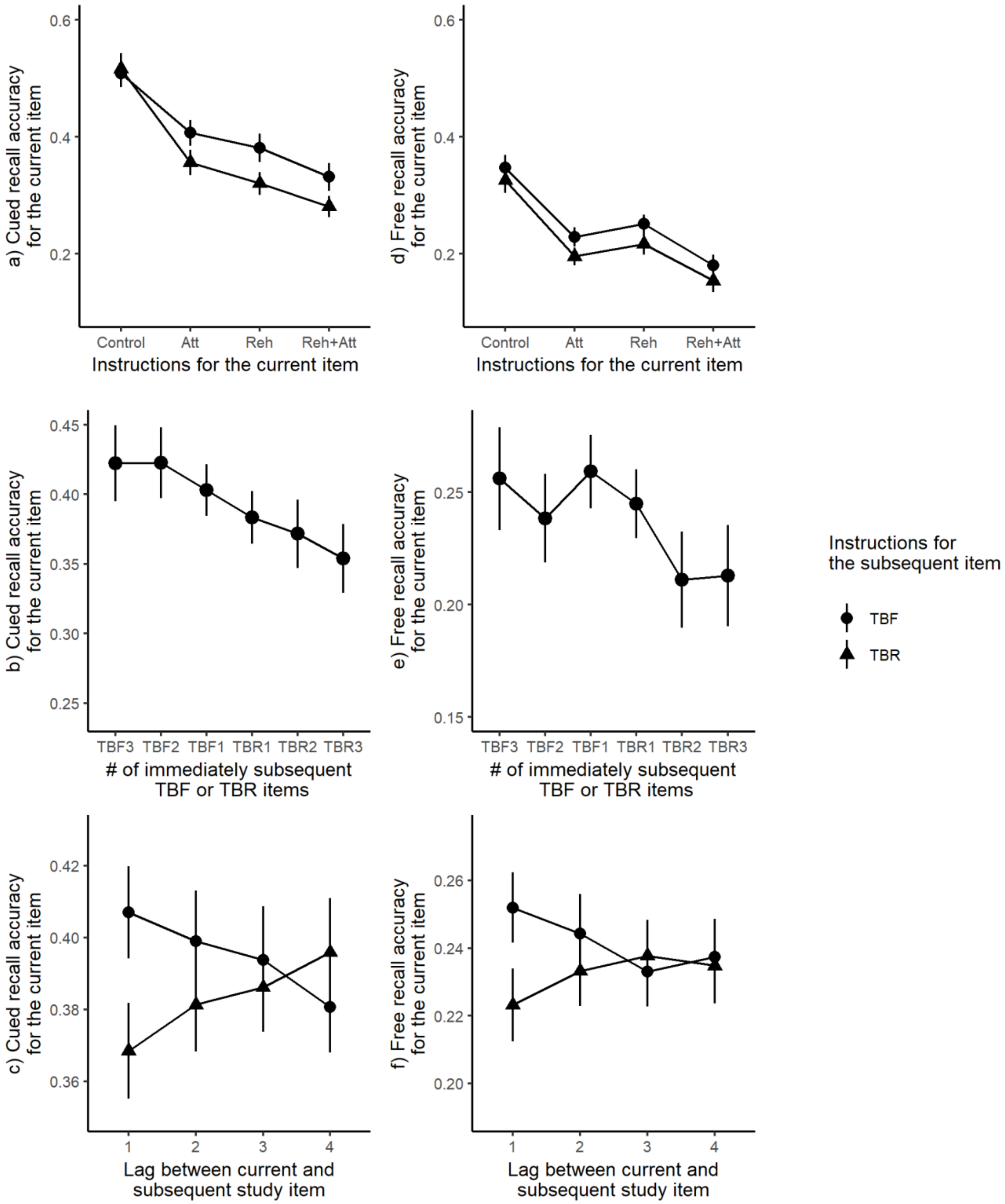

Figure S2. *Results of Experiment 2 – cued recall (a,b,c) and free recall (d,e,f) for the current item depending on a, d) whether it was* ***followed*** *during study by a TBR or a TBF item and the dual task condition (Control = No dual task, Att = Divided attention, Reh = suppressed rehearsal, Reh+Att = simultaneous divided attention and suppressed rehearsal; b, e) how many of the immediately* ***following*** *items during study were TBR or TBF; c, f) what was the study position lag between the current and the* ***subsequent*** *item. Error bars represent ±1 SE.*

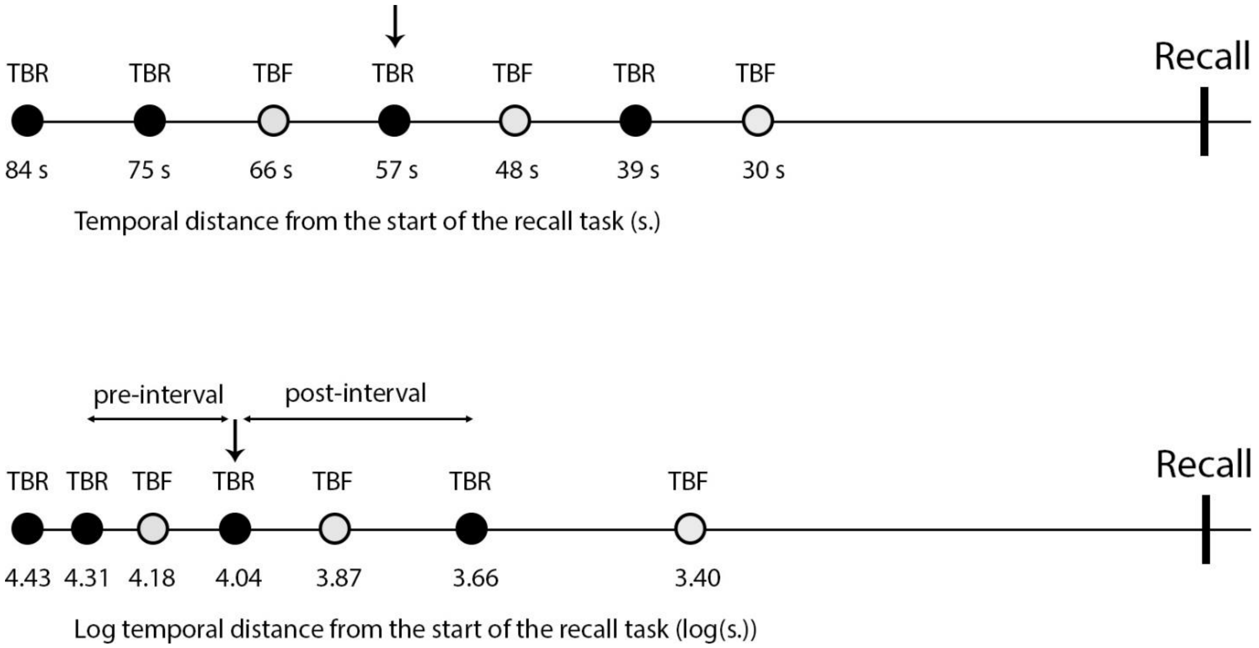

Figure S3. Illustration of SIMPLE's time compression and effects of temporal distinctiveness within the DF task. The top panel represents the study sequence in real time. The arrow indicates the item of interest that is preceded and followed by TBF items. The bottom panel represents the logarithmic compression representation in SIMPLE. If TBF items are not stored a preceding TBF item will create a smaller temporal isolation gap (the pre-interval) compared to the gap created by the subsequent TBF item (the post-interval).

B. Attentional refreshing versus resource depletion

Our claim is that the attentional suppression task would not remove a resource depletion effect but that it would remove any effect due to attentional refreshing. We consider it crucial to consider *when* resources deplete. The resource depletion account postulates that processing of TBR-items depletes more resources than processing of TBF-items at the very time these items are actually processed. In contrast, the attentional refreshing account assumes that processing of both TBR and TBF items initially depletes the same amount of resources but that, when the subsequent word n is processed, resources are split between item n and item n-1 if n-1 was TBR but not if it was TBF.

Why does the timing of processing preceding items matter? Because we posit that the effect of dividing attention in this task is two-fold. First, it reduces the total amount of available resources. Second, it renders it less likely that participants use their remaining resources to refresh preceding items, that is, they are expected to focus on the current item only (just as with the rehearsal suppression task). The reduction of overall resources should not remove the DF after-effect. If anything, it should increase it, as we have demonstrated empirically elsewhere for other item strength effects that are accounted for by the same depletion/recovery principle (Popov & Reder, 2018). The size of the effect depends on the degree of the working memory load imposed by the secondary task, with heavier loads causing stronger effects (Shen, Popov, Delahay & Reder, 2018). In other words, the resource depletion account predicts that divided attention will hurt overall performance but that the benefit for a particular item from being preceded by a TBF item remains (or is enhanced), because preceding TBR items still consume more resources than TBF items. In contrast, the attentional refreshing account would predict that, with divided attention, the tendency to refresh either preceding item (TBF or TBR) is eliminated (reduced), which means that the advantage of preceding TBF relative to TBR items is absent (or diminished in case the secondary task would not be attention demanding enough).

Figures S4-6 illustrate this explanation. Figure S4 illustrates our preferred resource depletion account (please also refer to the corresponding captions for a detailed description). It also illustrates why redirecting attention to a secondary task would not moderate the DF after effect – having fewer overall resources does not alter the differential depletion effect produced by TBR and TBF items. In contrast, Figures S5 and S6 illustrate the alternative attentional refreshing account. Figure S5 illustrates how attentional refreshing would account for the DF after-effect in Experiment 1, and the control condition in Experiment 2, where attention was not divided. The main difference between Figure S4 and Figure S5 concerns the difference between proactive and retroactive depletion. According to the resource depletion account, the effect of depleting more resources for TBR n-1 items is proactive and occurs in panel b/e, at the time at which the item is studied. According to the attentional refreshing account, the effect of depleting more resource for TBR n-1 items is retroactive, and it occurs in panel c/f, at the time at which the subsequent item n is being processed (i.e., participants redirect some of their attention backwards to previous TBR items). Figure S6 illustrates why the attentional refreshing account would not predict a DF after-effect when attention is divided. The lighter, narrower, dashed arrows indicate that participants are less likely to redirect their attention to previous items, because their resources are more limited. This should reduce the likelihood of refreshing a preceding TBR-item to a significant extent, rendering its processing more similar to a preceding TBF-item. We believe this is a reasonable formulation of the attentional refreshing account, since when faced with distractions, it is more rational to prioritize the processing of incoming items relative to already processed items.

An illustration of the resource depletion account of DF after-effects

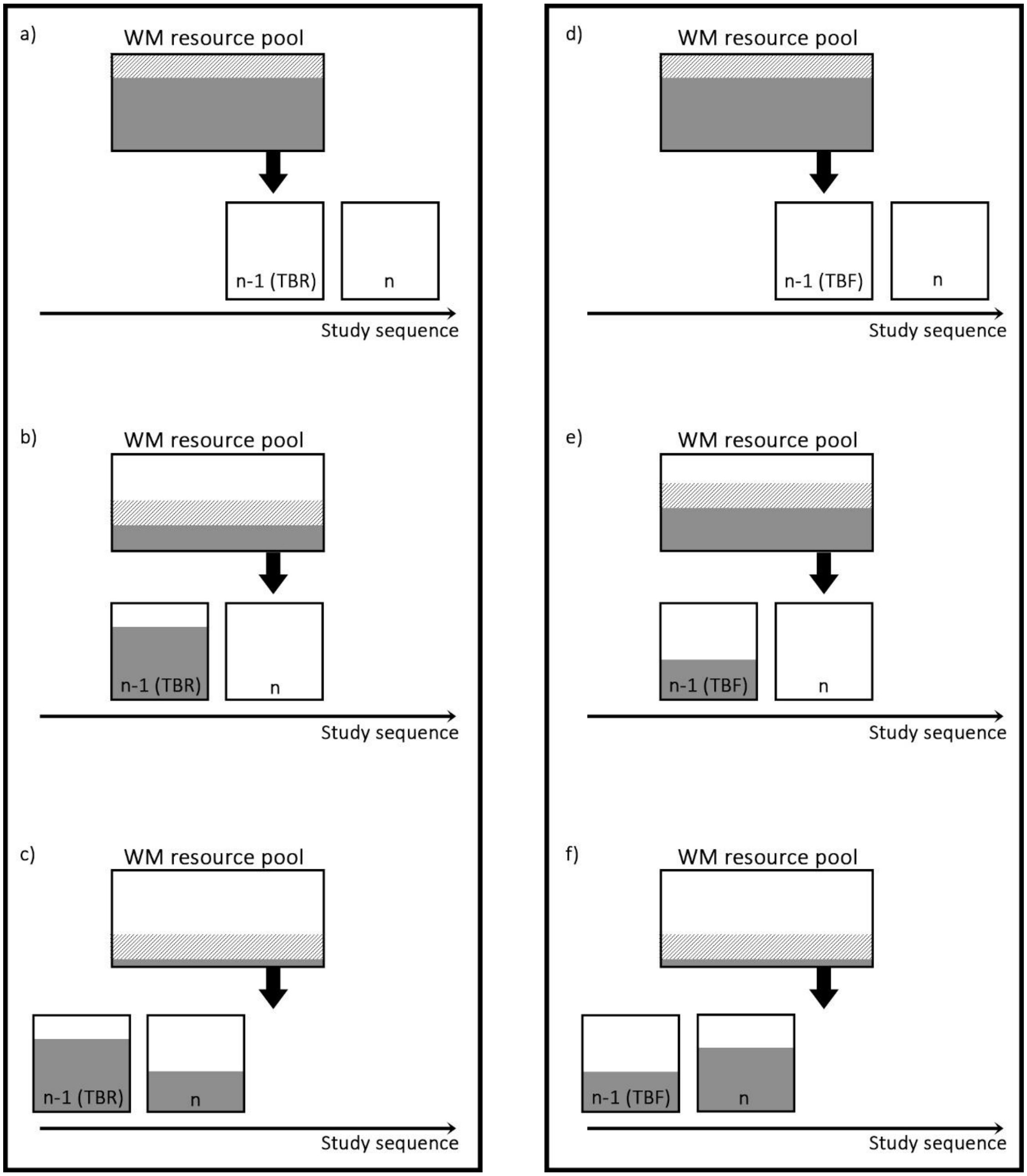

Figure S4. An illustration of how dividing attention in the directed forgetting task is reflected by the resource depletion account. Left panels (a,b,c) illustrate the case where the *n-1* item in the study sequence is a to-be-remembered (TBR) item. Right panels (d,e,f) illustrate the case where the *n-1* item in the study sequence is to-be-forgotten. Within each panel, the top box reflects the working-memory (WM) resource pool. The dark-shaded area within the pool is the currently available WM resource, while the stripped area reflects the part of WM that is devoted to the monitoring the secondary parity task (i.e., unavailable resources due to attentional suppression. The square boxes below the WM pool labeled "n-1 (TBR)" and "n" are the study items, and the dark shaded area reflects their strength. The down-turned arrows reflect where the WM resource is deposited. The study sequence begins at panel a), when item n-1 appears. After participants study it, its strength increases and the WM pool is partially depleted (b), when the study item *n* appears. If item *n-1* was TBR (a,b,c), rather than TBF (d,e,f), there are fewer WM resources available for processing item n (e.g. less dark shaded area in WM pool in b vs e). This results in item n being weaker if preceded by a TBR n-1 item (c), rather than a TBF n-1 item (f). The attentional suppression task does not affect the DF after-effect.

An illustration of the attentional refreshing account of DF after-effects

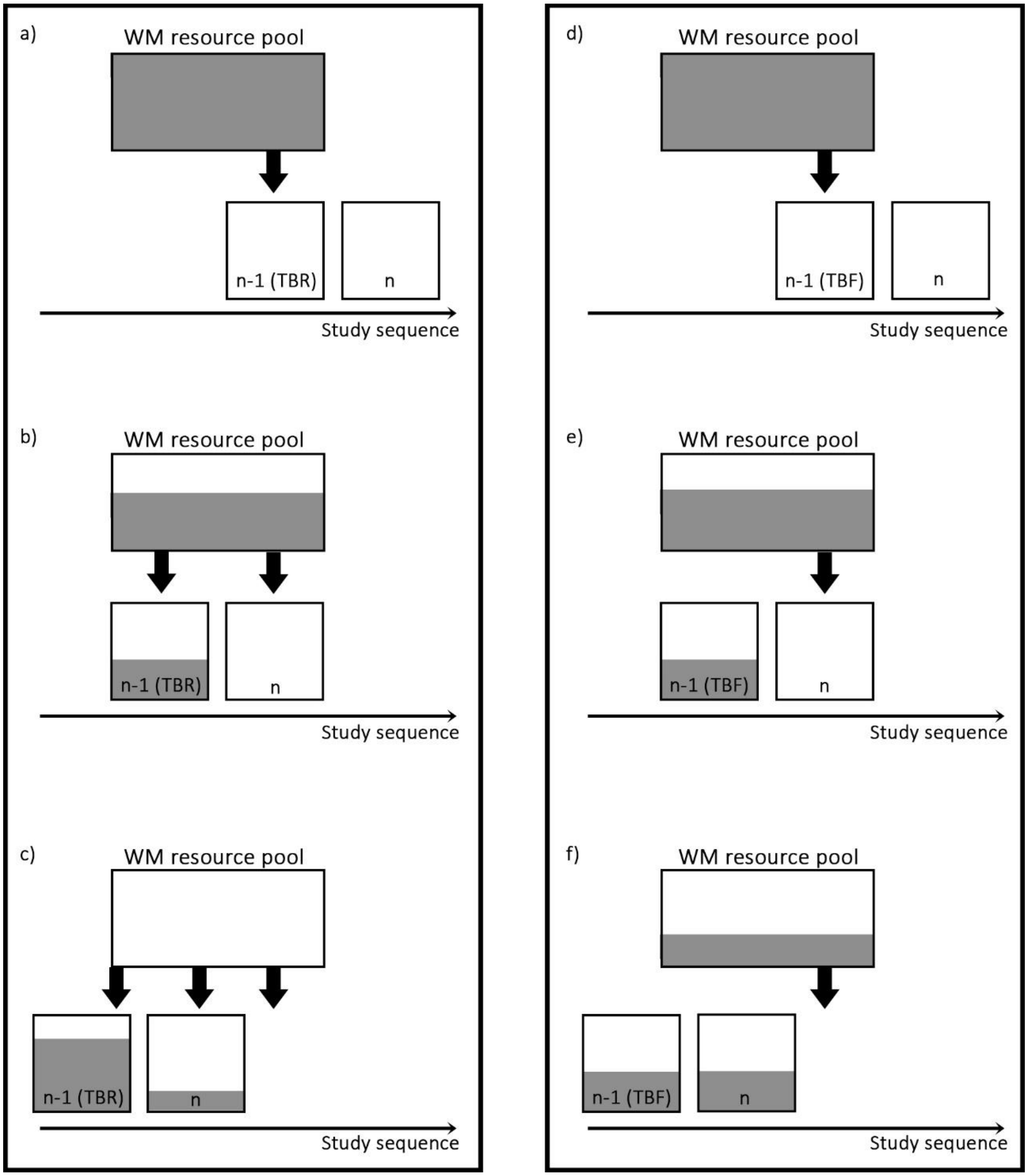

Figure S5. An illustration of DF after-effects according to the alternative attentional refreshing account. Left panels (a,b,c) illustrate the case where the *n-1* item in the study sequence is a to-be-remembered (TBR) item. Right panels (d,e,f) illustrate the case where the *n-1* item in the study sequence is to-be-forgotten. Within each panel, the top box reflects the WM resource pool. The dark-shaded area within the pool is the currently available WM resource, while the stripped area reflects the part of WM that is devoted to the monitoring the secondary parity task (i.e., unavailable resources due to attentional suppression). The square boxes below the WM pool labeled "n-1 (TBR)" and "n" are the study items, and the dark shaded area reflects their strength. The down-turned arrows reflect where the WM resource is deposited. The study sequence begins at panel a), when item n-1 appears. After participants study it, its strength increases and the WM pool is partially depleted (b), when the study item *n* appears. If item *n-1* is TBR (a,b,c), rather than TBF (d,e,f), participants will redirect some of their attention back to the preceding item (n-1) while processing the current item *n* (as reflected by the two black arrows in panel b vs e). This results in item n being weaker if preceded by a TBR n-1 item (c), rather than a TBF n-1 item (f).

An illustration of attentional suppression removing the DF after effect in the attentional refreshing account

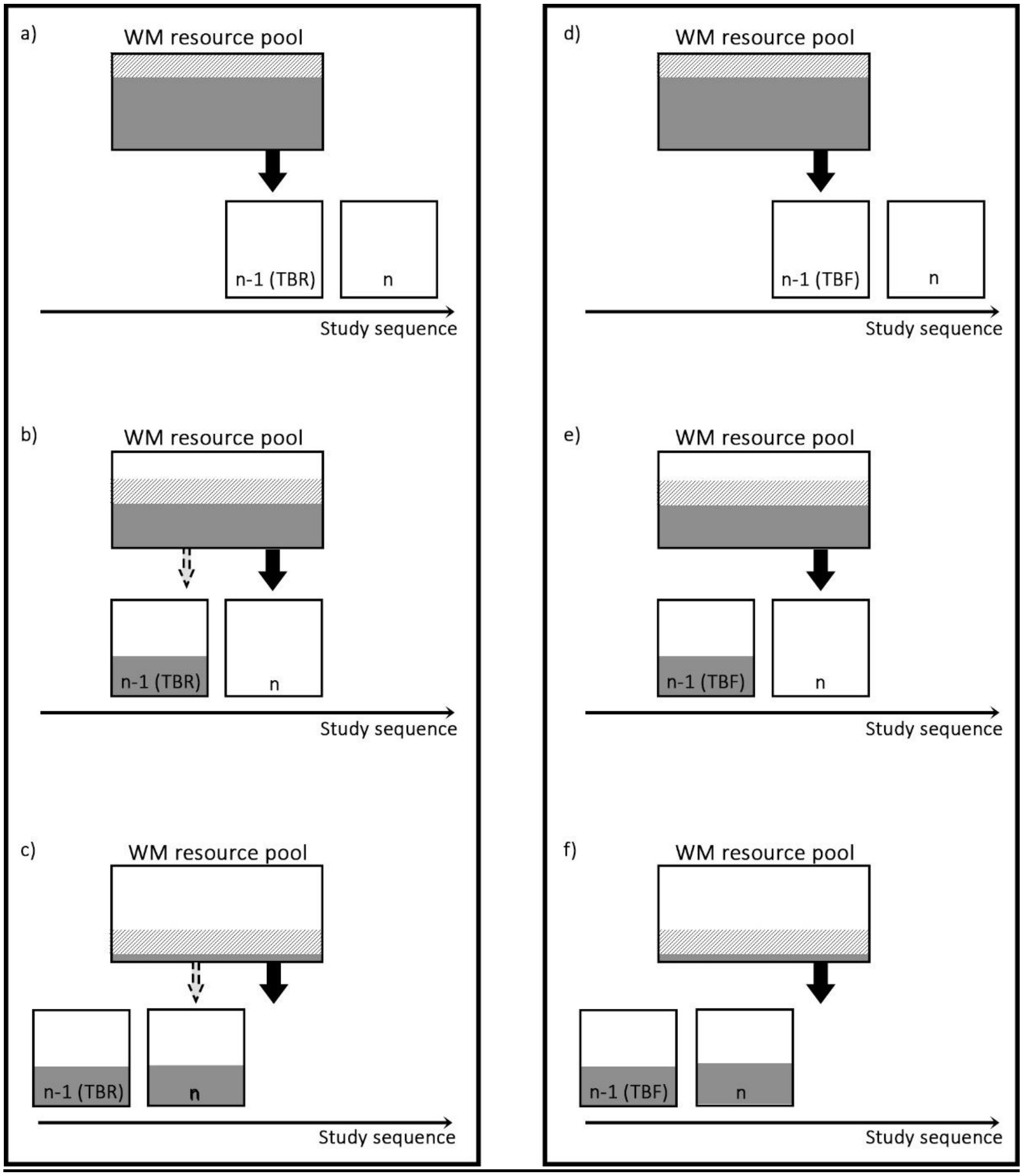

Figure S6. An illustration of why the attentional suppression secondary task should remove the DF after effect in the attentional refreshing account. See figure S4 for the main description. When attention is suppressed by the secondary task, participants are less likely to redirect their attention back to previously encountered items, as reflected in the slimmer, lighter, dashed arrows in b) and c). As a result, in contrast to the situation illustrated in Figure S4, whether the preceding item was TBR or TBF does not differentially affect the strength of item n.

C. *Detailed description of the SAC model*

The theory we present is an evolution of the Source of Activation Confusion model (Reder et al., 2000; 2007a&b; Reder & Schunn, 1996; Schunn, Reder, Nhouyvanisvong, Richards, & Stroffolino, 1997), which itself has roots in the ACT-R cognitive architecture (Anderson et al., 2004). The model has been successful in accounting for a variety of findings, in particular in recognition and cued-recall memory, including the key mirror frequency (Reder et al., 2000), list length (Cary & Reder, 2003) and list strength effects (Diana & Reder, 2005). For that reason, we have imported many of its assumptions in the current theory. *This is a condensed description of the model* containing only the information relevant for modeling the current studies. For a complete description of the SAC model, its deviations from previous versions (e.g., Reder et al. 2000; 2007), and its application to a variety of different tasks and manipulations, see Popov & Reder (2018).

1. *Representation.*

Memory traces are represented as interconnected nodes in an associative network. There are three types of nodes: for concepts (e.g., the concepts representing each word in a word pair), for episodes ("I studied the word pair *chair-apple* in this experiment"), and for contextual information (the internal and external context associated with an episode). Figure S7 shows a basic schematic illustration of the model representations in the current paired-associate list learning memory studies. Episodic nodes link together the individual aspects of experiences.

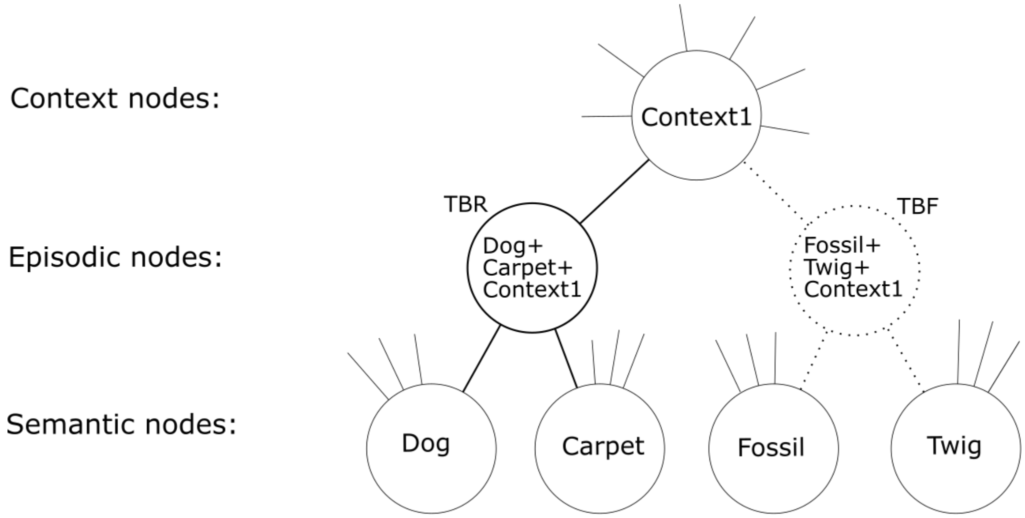

Figure S7. *Illustration of the SAC model structure for the current item-based directed forgetting experiments. Participants studied the word pairs Dog-Carpet and Fossil-Twig (among others) and were given instructions to remember Dog-Carpet (TBR), but to forget Fossil-Twig (TBF). Each concept has a pre-existing semantic node, which has connections to multiple episodes in which it has been experienced over time. The current list context also has a separate node, which is connected to all the episodes (i.e. different trials) experienced in the current list. There is a unique episode node that connects all features of an experience, i.e. the two concepts and the context in which they are experienced. TBR nodes are created with a greater base-level strength than TBF nodes, as reflected in the thicker border line of the TBR node.*

2. Learning, forgetting, and base-level strength.

Each node in memory has two important properties, namely base-level strength and current activation (Reder et al, 2000). These values are a function of experience and differ between nodes. When a word pair is studied, the nodes representing the individual words are activated, an episode node connecting them is created and their base-level strength is increased by a discrete amount. These increments in base-level strength decay over time. The increase in base-level strength depends on node strength at the time of study, which is the first deviation from the original version of SAC. Specifically, nodes can reach a maximum strength level of 1, and each increment strengthens the node as a proportion δ (learning rate) of the maximum strength minus its current base-level strength, *B*:

$$s = \Delta B = \delta(1 - B) \tag{1}$$

The size of the increment calculated by Equation 1 is the default, assuming there are sufficient resources available; if WM resources are insufficient, this increment is modified by Eq. 5. We initialize all new nodes with a base-level strength of δ, because they have no prior strength.

In the DF paradigm, an instruction to *forget* a word pair does not change the base-level strength of the initially created node. However, the instruction to *remember* a pair further strengthens the node according to the same equation. For example, if the learning rate is δ = 0.553 (the actual estimate for Experiment 1), we create an initial node with strength of 0.553 regardless of the instruction type. Whenever a remember instruction appears, the corresponding node strength is increased additionally by 0.553 * (1-0.553) = 0.247, and the resulting node strength is 0.553+0.247 = 0.8.

This strengthening equation has several benefits (see Popov & Reder, 2018 for a discussion). Most importantly, the working memory cost of an increment of size s can be set to be proportional to s. Since weaker items are strengthened more, i.e., $\delta(1 - B_{weak}) > \delta(1 - B_{strong})$, their strengthening resource cost is also larger.

The main cause of forgetting in the model is that base-level strength decays over time. Each strength increment decays independently[1] depending on how much time has passed since its occurrence. Thus, at any time *t*, the base-level strength of a node is:

$$B = B_0 + \sum_{i=1}^{n-1} s_i \times (1 + t - t_i)^{-d}, \tag{2}$$

where s_i is the strength increment produced by the *i*-th repetition, $t - t_i$ is the time that elapsed since the *i*-th repetition, *d* is the decay rate, and B_0 is the preexisting base-level strength. The initial time value is offset by 1, so that immediately after encoding the increment size is not infinite.

The links that connect individual nodes also vary in strength, depending on how often the two nodes have been co-active. The increment and decay of link strength also follow Equations 1 and 2 and the only difference is in the values of the decay parameter.

[1] The independent decay of each increment is not relevant for the current study, since each item was studied only once, and the TBR instruction causes an additional increment at the same time as the node creation. For more information on why we chose such a function, see Popov & Reder, 2018.

3. Strengthening and binding deplete working memory resources.

The key novel aspect of the model is that learning is fueled by a shared pool of resources, and that greater increments cost more resources. We assume that people have different total amount of WM resources, which is denoted by a W_{max} parameter. Every time a node/link is created, it is strengthened by an amount s, and s^2 amount of resources is depleted. Under most circumstances, this defaults to Equation 1, squared:

$$W_{default\_cost} = s^2 = \left(\delta(1-B)\right)^2 \tag{3}$$

Since TBR items are incremented more overall, their processing depletes more resources than TBF items. We chose the cost of strengthening to be s^2 because the square exponent slightly increases the cost difference between small and big increments relative to the overall cost of the operations, which led to better fits of most models presented in Popov & Reder (2018).

We also assume that the resource pool replenishes at a linear function of time since the last operation, $t - t_i$, and the remaining resources at time t_i, such that:

$$W_t = \min\left(W_{max},\, W_{t_i} + \, w_r(t - t_i)\right) \tag{4}$$

where W_{t_i} is the amount of resource remaining after operation i and w_r is the recovery rate per second. Thus, after completing an operation, the resource pool recovers at a fixed rate until it reaches W_{max}. WM depletion and recovery are illustrated in Figure S8, which shows the available resources at the beginning and end of each study trial in a single list that contains both TBF and TBR items. TBR items deplete more resources and the available resources are reduced when more items in a row are TBR.

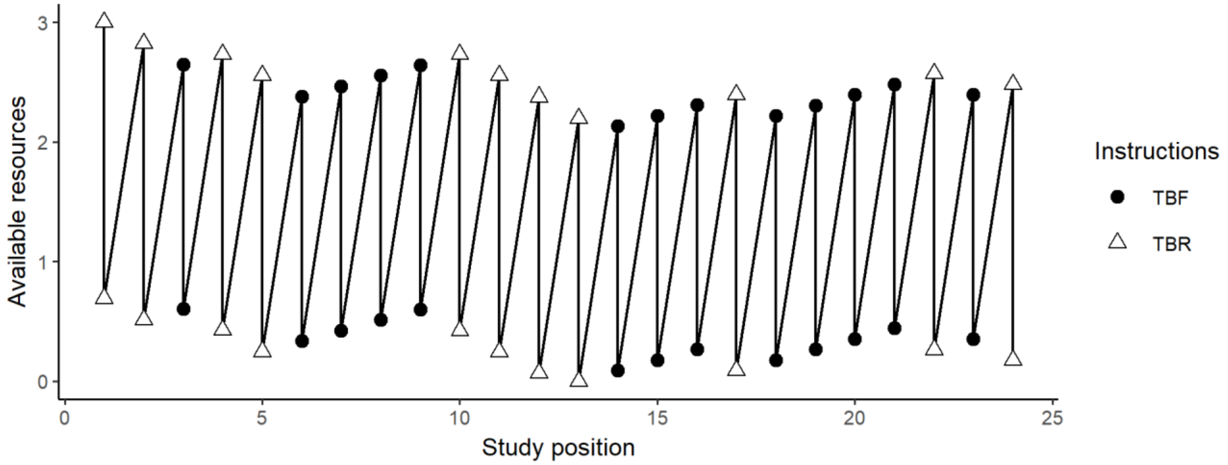

Figure S8*. Illustration of resource depletion and recovery in the model. Amount of available resources at the beginning and end of each trial during a single study list, as a function of item position on the list and whether the instructions were to remember (TBR) or to forget it (TBF).*

Finally, we need to describe the situation in which the remaining resources are less than the default cost of a process. We assume that the system uses whatever resource remains, and the strength increment in Equation 1 and 3 is reduced proportionally by $\sqrt{\frac{W_t}{W_{defaultcost}}}$:

$$s = \min\left(\sqrt{\frac{W_t}{W_{defaultcost}}}, 1\right) \times s =$$
$$\min\left(\sqrt{\frac{W_t}{{s_n}^2}}, 1\right) \times s =$$
$$\min(\sqrt{W_t}, s) =$$
$$\min\left(\sqrt{W_t}, \delta(1 - B)\right) \tag{5}$$

As a result, when the resources are insufficient, the memory trace strength is incremented by the square root of the remaining resources, $\sqrt{W_t}$ (the square root is due to the square exponent in Eq. 3).

4. Current activation and spreading activation.

Retrieval of nodes is based on their current level of activation. Nodes are activated directly when a concept is perceived, or indirectly by spreading activation from other nodes. The current activation decays exponentially, and its size is dependent on the node base level strength:

$$A = B \times A_{boost} \times e^{-\gamma t} \tag{6}$$

where B is the current base-level strength of the node, t is the time since activation, γ is the exponential decay parameter, and A_{boost} is the amount of activation received by the node. The current activation is thus a function of both the base-level strength and the size of the activation boost.

The size of the activation boost depends on whether the node is perceived directly, or whether activation spreads from other nodes. When a person perceives the word-pair “dog-carpet”, this activates not only the semantic nodes for “dog” and “carpet”, but also all event nodes to which these concepts are connected, as well as other concepts related to them. Following Reder et al. (2000), we assume that all nodes connected to a source of activation compete with each other:

$$A_{boost,r} = \sum_{s=1}^{n}\left(A_s \times \frac{S_{s,r}}{\sum_{i=1}^{k} S_{s,i}}\right) \tag{7}$$

where $A_{boost,r}$ is the boost in activation in the receiving node, A_s is the activation of the source node, $S_{s,r}$is the strength of the link between the source and the receiving node, and $\sum_{i=1}^{k} S_{s,i}$ is the summed strength of all links emanating from the source node. As a result, the more links a source node has and the stronger they are, the less activation is received by any specific node connected to the source (see Reder et al. 2007a & b, and Popov & Reder, 2018 for a discussion of the purpose behind this assumption).

5. Memory retrieval.

During a memory test, the network is cued by activating nodes representing the relevant cues, and by spreading activation from those nodes to all related nodes in the network. For cued recall, the cues are the list context, and the concept node for the cue word. For free recall, the only cue is the list context node. Additionally, for the free recall test we assume that there is output interference in that initial responses interfere with retrieving additional items, which we simulate by exponentiating the activation values. This results in squashing the activation of weak items compared to stronger items. In both cases, we assume that a response will be retrieved if the corresponding episode node's activation is above its retrieval threshold. Formally, we follow the signal detection theory tradition and assume that there is noise in the signal and that the probability of a response is the area to the right of a threshold under the normal distribution curve with a mean equal to the node's activation. Thus, if the node activation is equal to the retrieval threshold, the probability of a response is 50%.

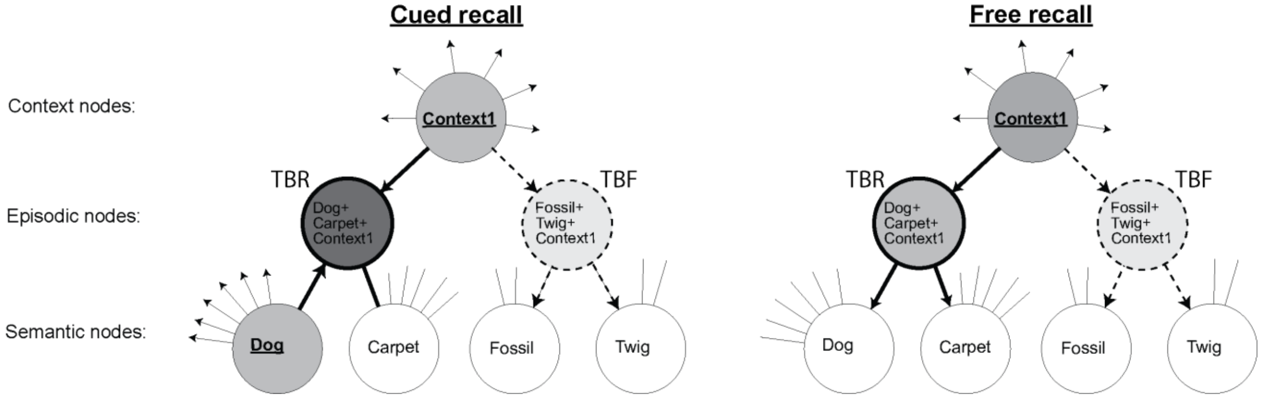

***Figure S9**. Illustration of spreading activation for cued and free recall (see Figure S6 for a description of node types). Participants study pair associates (e.g. dog-carpet), and are tested with either a 1) cued recall test (what word was associated with "apple"?) or 2) free recall test (recall all words presented in the previous list). While the contents of memory is the same, the amount of activation reaching the episode nodes differs depending on which cues are presented. Underlined text represents the cues for each retrieval task (i.e., context1 and dog). The shade darkness of nodes represents their activation levels. The activation is higher for TBR than for TBF items.*

6. *Modeling details.*

SAC is a process model that takes the sequence of trials performed by each participant and applies each operation on a trial-by-trial basis. This results in activation values for each test trial, which we convert into response probabilities as described in the preceding section. We fit the model by generating predicted response probabilities through a simulation for each trial and each participant, then summarizing the response probabilities over all subjects and separately for each condition of interest. Six parameters (learning rate and resource recovery rate, as well as the retrieval thresholds and activation noise for free and cued recall) were optimized by minimizing the root mean squared error of the cued recall and free recall data averaged over all subjects, the current instruction type and the number of consecutive preceding TBR or TBF items. The optimization was performed using the downhill simplex algorithm as implemented in Python's Scipy library. All remaining parameters had the default values we have used in prior models (see Popov & Reder, 2018). All parameter values are summarized in Table S1.

Table S1 Description of SAC parameters

Parameter	**Description**	**Exp. 1**	**Exp. 2**
d_n	Power decay rate for node base-level strength	-0.180	-0.180
d_l	Power decay rate for link strength	-0.120	-0.120
y	Exponential decay rate for current activation	0.200	0.200
δ	Learning rate for base-level strength	**0.553**	**0.639**
W_{max}	Total WM resource capacity	3.000	3.00
w_r	WM recovery rate	**0.526**	**0.551**
θ_{cued}	Cued recall retrieval threshold for episodic nodes	**0.219**	**0.279**
σ_{cued}	Cued recall standard deviation of the noise added to episodic activation	**0.831**	**0.451**
θ_{free}	Free recall retrieval threshold for episodic nodes	**0.167**	**0.457**
σ_{free}	Cued recall standard deviation of the noise added to episodic activation	**0.431**	**0.868**

Note: **bold-underlined** parameters were free to vary in estimating the model. The remaining parameters were fixed and imported from other SAC models

VIII. References

Basden, B. H., Basden, D. R., & Gargano, G. J. (1993). Directed forgetting in implicit and explicit memory tests: A comparison of methods. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *19*(3), 603–616.

Bjork, R. A. (1970). Positive forgetting: The noninterference of items intentionally forgotten. *Journal of Verbal Learning and Verbal Behavior*, *9*(3), 255–268.

Bjork, R. A. (1972). Theoretical implications of directed forgetting. In A. W. Melton & E. Martin (Eds.), *Coding processes in human memory* (pp. 217–235). Washington, DC: Winston.

Brown, G. D., Neath, I., & Chater, N. (2007). A temporal ratio model of memory. *Psychological Review*, *114*(3), 539.

Bürkner, P.-C. (2017). brms: An *R* package for Bayesian multilevel models using *Stan*. *Journal of Statistical Software*, *80*(1).

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.

Cary, M., & Reder, L. M. (2003). A dual-process account of the list-length and strength-based mirror effects in recognition. *Journal of Memory and Language*, *49*(2), 231–248.

Castel, A. D., & Craik, F. I. M. (2003). The effects of aging and divided attention on memory for item and associative information. *Psychology and Aging*, *18*(4), 873–885.

Craik, F. I., Govoni, R., Naveh-Benjamin, M., & Anderson, N. D. (1996). The effects of divided attention on encoding and retrieval processes in human memory. *Journal of Experimental Psychology. General*, *125*(2), 159–180.

Criss, A. H., Aue, W. R., & Smith, L. (2011). The effects of word frequency and context variability in cued recall. *Journal of Memory and Language*, *64*(2), 119–132.

Detterman, D. K. (1975). The von Restorff effect and induced amnesia: Production by manipulation of sound intensity. *Journal of Experimental Psychology: Human Learning and Memory*, *1*(5), 614.

Diana, R. A., & Reder, L. M. (2005). The list strength effect: A contextual competition account. *Memory & Cognition*, *33*(7), 1289–1302.

Diana, R. A., & Reder, L. M. (2006). The low-frequency encoding disadvantage: Word frequency affects processing demands. *Journal of Experimental Psychology: Learning Memory and Cognition*, *32*(4), 805–815.

Epstein, W. (1972). Mechanisms of directed forgetting. *Psychology of Learning and Motivation - Advances in Research and Theory*, *6*(C), 147–191.

Fawcett, J. M., & Taylor, T. L. (2008). Forgetting is effortful: Evidence from reaction time probes in an item-method directed forgetting task. *Memory & Cognition*, *36*, 1168–1181

Fawcett, J. M., & Taylor, T. L. (2012). The control of working memory resources in intentional forgetting: Evidence from incidental probe word recognition. *Acta Psychologica*, *139*(1), 84-90.

Gelman, A., & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, *7*, 457–472.

Golding, J. M., & MacLeod, C. M. (1998). *Intentional forgetting: Interdisciplinary approaches*. Mahwah, NJ US: Erlbaum.

Heister, J., Würzner, K.-M., Bubenzer, J., Pohl, E., Hanneforth, T., Geyken, A., & Kliegl, R. (2011). dlexDB-eine lexikalische Datenbank für die psychologische und linguistische Forschung. *Psychologische Rundschau*, *62*(1), 10–20.

Hourihan, K. L., Ozubko, J. D., & MacLeod, C. M. (2009). Directed forgetting of visual symbols: Evidence for nonverbal selective rehearsal. *Memory & Cognition*, *37*(8), 1059-1068.

Hulme, C., Stuart, G., Brown, G. D. ., & Morin, C. (2003). High- and low-frequency words are recalled equally well in alternating lists: Evidence for associative effects in serial recall. *Journal of Memory and Language*, *49*(4), 500–518.

Kamp, S. M., Lehman, M., Malmberg, K. J., & Donchin, E. (2016). A Buffer Model account of behavioral and ERP patterns in the Von Restorff paradigm. *AIMS Neuroscience*, *3*(2), 181-202.

Lee, M. D., & Wagenmakers, E.-J. (2013). *Bayesian cognitive modeling: A practical course*. Cambridge: Cambridge University Press.

Lehman, M., & Malmberg, K. J. (2013). A buffer model of memory encoding and temporal correlations in retrieval. *Psychological Review*, *120*(1), 155.

Malmberg, K. J., & Murnane, K. (2002). List composition and the word-frequency effect for recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *28*(4), 616–630.

Malmberg, K. J., & Nelson, T. O. (2003). The word frequency effect for recognition memory and the elevated-attention hypothesis. *Memory & Cognition*, *31*(1), 35–43.

Marevic, I., Arnold, N. R., & Rummel, J. (2018). Item-method directed forgetting and working memory capacity: A hierarchical multinomial modeling approach. *The Quarterly Journal of Experimental Psychology*, *71*(5), 1070-1080.

McFarlane, K. A., & Humphreys, M. S. (2012). Maintenance rehearsal: The key to the role attention plays in storage and forgetting. *Journal of Experimental Psychology: Learning Memory and Cognition*, *38*(4), 1001–1018.

Pastötter, B., Tempel, T., & Bäuml, K.-H. T. (2017). Long-term memory updating: The reset-of-encoding hypothesis in list-method directed forgetting. *Frontiers in Psychology, 8*.

Plummer, M. (2003). JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. *Proceedings of the 3rd International Workshop on Distributed Statistical Computing*, (Dsc), 1–10.

Popov, V., & Reder, L. (2018). Frequency effects on memory: A resource-limited theory. *PsyArXiv*. Retrieved from http://doi.org/10.17605/OSF.IO/DSX6Y

Reder, L. M., Liu, X. L., Keinath, A., & Popov, V. (2016). Building knowledge requires bricks, not sand: The critical role of familiar constituents in learning. *Psychonomic Bulletin and Review*, *23*(1), 271–277.

Reder, L. M., Nhouyvanisvong, A., Schunn, C. D., Ayers, M. S., Angstadt, P., & Hiraki, K. (2000). A mechanistic account of the mirror effect for word frequency: A computational model of remember–know judgments in a continuous recognition paradigm. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *26*(2), 294.

Reder, L. M., Paynter, C., Diana, R. A., Ngiam, J., & Dickison, D. (2007). Experience is a double-edged sword: A computational model of The encoding/retrieval trade-off with familiarity. In *Psychology of Learning and Motivation - Advances in Research and Theory* (Vol. 48, pp. 271–312). Elsevier.

Reder, L. M., Oates, J.M., Dickison, D., Anderson, J.R., Gyulai, F., Quinlan, J.J., Ferris, J.L., Dulik, M. & Jefferson, B. (2007b). Retrograde facilitation under midazolam: The role of general and specific interference. *Psychonomic Bulletin & Review, 14*(2), 261-269.

Rummel, J., Marevic, I., & Kuhlmann, B. G. (2016). Investigating storage and retrieval processes of directed forgetting: A model-based approach. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *42*(10), 1526–1543.

Sahakyan, L., & Kelley, C. M. (2002). A contextual change account of the directed forgetting effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 28*(6), 1064-1072

Schunn, C. D., Reder, L. M., Nhouyvanisvong, A., Richards, D. R., & Stroffolino, P. J. (1997). To calculate or not to calculate: A source activation confusion model of problem familiarity's role in strategy selection. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *23*(1), 3.

Sederberg, P. B., Howard, M. W., & Kahana, M. J. (2008). A context-based theory of recency and contiguity in free recall. *Psychological Review*, *115*(4), 893–912.

Shen, Z., Popov, V., Delahay, A. B., & Reder, L. M. (2018). Item strength affects working memory capacity. *Memory and Cognition*, *46*(2), 204–215.

Unsworth, N., & Spillers, G. J. (2010). Variation in working memory capacity and episodic recall: The contributions of strategic encoding and contextual retrieval. *Psychonomic Bulletin & Review*, *17*(2), 200–205.

Wagenmakers, E.-J., Morey, R. D., & Lee, M. D. (2016). Bayesian benefits for the pragmatic researcher. *Current Directions in Psychological Science*, *25*(3), 169–176.

Watkins, M. J., LeCompte, D. C., & Kim, K. (2000). Role of study strategy in recall of mixed lists of common and rare words. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *26*(1), 239.