Accessibility of Working Memory Representations in the Focus of Attention: Heightened or Reduced?

Hautekiet, Caro Béatrice G; Langerock, Naomi; Vergauwe, Evie

Supplementary material A

Experiment 1

A wordlist of 484 French words necessary for the experiment (16 for training and 468 for the actual experiment) was created using the Lexique 3 database (New, 2006). Within the database, a set of nouns was selected consisting of six letters, two syllables, four to six phonemes, and a book frequency of two to 50. Plural nouns were taken out as well as ambiguous words, words with an obvious negative connotation, words with double significations, words derived from English words, words pronounced the same as another word of the wordlist, and derivations of verbs. From the remaining words, the most frequent words were selected for the experimental trials, and the following 16 most frequent words were used for the practice trials. One hundred-eleven nonwords were selected from a large database containing French nonwords (Ferrand et al., 2010) based on the criterion that the nonword should consist of six letters and two syllables, as for the actual words. Three of these nonwords were used for the instructions and practice trials.

To counterbalance the use of the different words in the different conditions, we created 26 different versions of the experiment. Each participant received one of these 26 versions (version 26 was not included in the final dataset since the participant that was assigned this version did not give permission for data usage). In 36 out of 216 total trials, a novel probe was presented. We made sure every word was used as a novel probe in two of the 26 versions. To do so, 13 different wordlists, consisting of 36 words matched for word frequency, were created. Each of these lists was used as the novel probes in two versions of the experiment and the remaining 432 words were used to make the word pairs presented at the beginning of each trial. The order of the word pairs was fixed in each version and corresponded to the first word of each of the remaining 12 lists (the 13^{th} list was used for the novel probes), then the second word of the remaining 12 lists, and so on. In the counterpart version (using the same list for the novel

probes), the position of the words (above or below the center of the screen) was switched around. The order of the arrows, pointing to one of the previously presented words, was randomly predetermined (108 in each direction) and remained the same throughout all versions of the experiment. In this way, each word was refreshed as often over the different versions of the experiment. The probe type used in each trial was randomly chosen by E-prime during the experiment. The order of the novel probes was predetermined and corresponded to the first word of the selected list until the 36th word. Proceeding in this way allowed us, first of all, to construct a list of the to-be-refreshed words in advance, to make sure the experimenter could easily note if the participant spoke aloud the correct word or not. Secondly, as the order of the words was in function of their frequency (starting with the highest frequent words), the frequency of the two words of the word pair and the probe used in the trial was similar and could hence not affect the RT for the words in the lexical decision task.

Experiment 2

In all 270 trials, two memory items were presented (i.e., 540 in total) and for every experiment, 60 novel probes were necessary. Additionally, 16 words were necessary for the instructions and practice trials. Thus, a wordlist of 616 French words necessary for the experiment was created using the following procedure. Within the Lexique 3 database (New, 2006), a set of nouns was selected all consisting of six letters, two syllables, four to six phonemes, and a frequency (book frequency) between 1 and 100. This resulted in a list of 873 words. After applying the same exclusion criteria as used in Experiment 1, we were left with 609 words exactly. From this list, 600 words were used for the actual experiments. For the 16 words necessary for the practice trials, we could use the remaining nine words and, by expanding our previously set criteria, another set of seven words with three phonemes were added. Ninety-two nonwords were selected from a large database containing French nonwords (Ferrand et al., 2010) based on the same criteria as used in Experiment 1. Two of these

nonwords were used in the practice trials and the other 90 were used in the actual experiment (30 in each CTOA condition).

In order to counterbalance the use of the different words in different conditions, 20 different versions of the experiment were created. Each participant received one of these 20 versions. In total there were 270 trials, of which 60 used a novel probe. We made sure every word was used as novel-probe in two of the 20 different versions. To do so, 10 different wordlists of 60 words were created, that were matched for word frequency (the 10 most frequent words were randomly distributed over the 10 lists, and the same was done for the next groups of 10 following most frequent words). Each of these lists was used as the novel probes in two versions of the experiment and the remaining 540 words were used to make the word pairs presented at the beginning of each trial. The order of the arrows was randomly predetermined (135 in each direction) and remained the same in all versions of the experiment.

The probe type, used for each trial, was randomly chosen by E-prime during the experiment. The order of the words to form the word pairs was fixed in each version and corresponded to the first word of each of the 9 lists (the tenth list was used for the novel probes), then the second word of each of the 9 lists, and so on. All other randomizations remained the same as used in Experiment 1.

Experiment 3

For this experiment, 218 French words were necessary (of which 18 were for training). Therefore, we derived a wordlist of 200 words from the existing wordlist of 600 French words used in Experiment 2 (originally from the Lexique 3 database; New, 2006). We made sure there was an equal distribution of words coming from different book frequencies to keep the same frequency range of the previous experiment (i.e., book frequency of 1 to 100). The 18 words necessary for training were randomly selected from the remaining words. Thirty-three nonwords (of which three for training) were randomly selected from the 90 nonwords used in

Experiment 2 (originally from Ferrand et al., 2010). The use of the different words was counterbalanced in the same way as was used in Experiment 2, resulting in 20 versions per CTOA condition. The order of the arrows was, again, randomly predetermined (45 in each direction) and remained the same throughout all versions of the experiment. The probe type, used for each trial, was randomly chosen by E-prime. All other randomizations and procedures were the same as used in Experiment 1.

Experiment 4a & 4b

The overall procedure of these experiments was the same as Experiment 1, except for the changes described in the method section of Experiments 4a and 4b. Additionally, to make sure the experimenter could follow whether the participant refreshed the correct word and responded in the correct way to the probe, we had to predetermine the triallists. More specifically, the order of the probe type (36 refreshed, 36 unrefreshed, 36 novel, and 108 nonwords) and direction of the arrows (108 in each direction) were randomly predetermined creating a fixed list. The order of the trials of this fixed list was reversed every two versions of the experiment (i.e., using the same novel words). Thus, as there were always two versions of the experiment using the same novel words (see also Experiment 1), for one of these two versions, we would use the predetermined list going from trial 1 to trial 216 and for the other version, we would use the predetermined list going from trial 216 to trial 1.

Supplementary material B

Experiment		Total trials excluded (%)	Total trials excluded (%) by probe type			
			Refreshed	*Unrefreshed*	*Novel*	*Nonword*
1		8	6	4	9	9
2	1600 ms	9	5	5	11	13
	2400 ms	8	5	4	12	11
	3200 ms	7	4	4	12	9
3	1600 ms	9	5	5	12	12
	3200 ms	8	4	4	13	10
4a		11	13	14	21	5
4b		8	10	9	13	5

Table. Following the preregistered exclusion criteria for each experiment, the proportion of excluded trials is presented in total and for trials including a refreshed, unrefreshed, novel, or nonword probe.

Supplementary material C

In Experiment 4a, we ceased data collection after testing 45 participants, although we had not reached the preregistered BF of 10 for or against one of our main t-tests, nor a max of N = 60. Indeed, as can be seen in Figure A, the sequential analysis of the second t-test, investigating a facilitative effect for an item in the focus of attention, showed that for the datasets included in the analysis, the BFs kept hovering around 10 between N = 25 and N = 36. With every batch of five additional participants, the BF always settled just below 10, although it exceeded the BF of 10 several times between participants 25 and 36 (after exclusion criteria). The sequential analysis of the first t-test, measuring an inhibition-of-return-like effect for an item in the focus of attention, was consistently inconclusive, overall shifting around a BF of 3 (see Figure B), thus it did not seem that this would become more definite with the addition of more participants. Therefore, we decided to deviate from our preregistration and terminate data collection after a total of 45 participants. It should be noted that optional stopping data collection is not considered a problem in Bayesian statistics (see e.g., Rouder, 2014). Additionally, it is important to note that data was collected during the COVID-19 pandemic. Since in-person testing had become more difficult, it was important not to waste valuable resources.

Reference

Rouder, J. N. (2014). *Optional stopping: No problem for Bayesians*. Psychonomic Bulletin & Review, 21(2), 301–308. 10.3758/s13423-014-0595-4

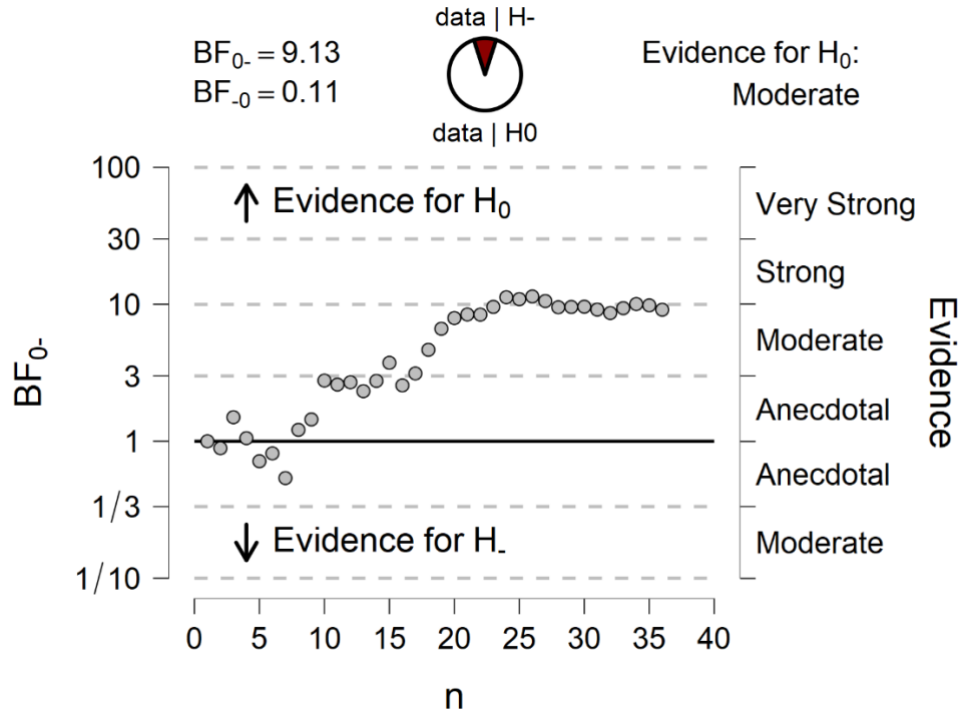

Figure A. Sequential analysis of the t-test investigating a facilitative effect in Experiment 4a.

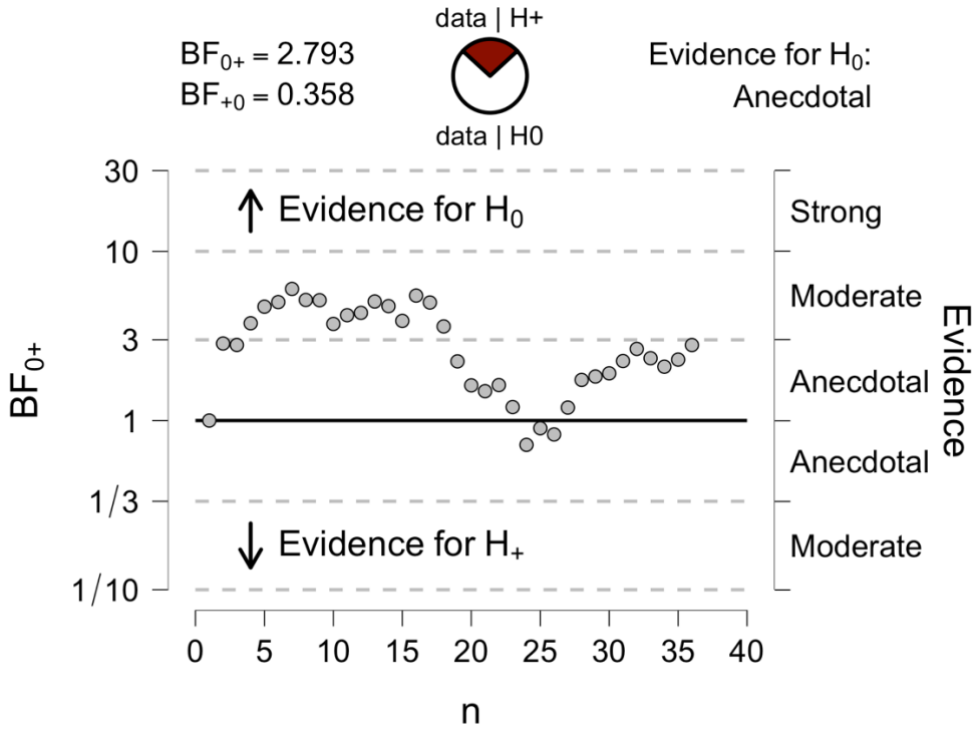

Figure B. Sequential analysis of the t-test investigating an inhibition-of-return-like effect in Experiment 4a.

Supplementary material D

In addition to the response modality hypothesis, we wanted to test a minor unpreregistered hypothesis in Experiment 4b. After studying the results of Experiments 1-4a, we ought it possible that the accessibility of an item in the focus of attention might depend on the length of the RTs to the presented probe because of a natural fluctuation of the accessibility of items in working memory, irrespective of the status within the focus of attention. More specifically, it might be possible that the reduced accessibility of an item in the focus of attention emerges when RTs to the presented probe are shorter, while it would become more accessible when RTs are prolonged. This differs from our timing hypothesis investigated in Experiments 2 and 3 as we were examining the time between when an item is brought into the focus of attention and when it is tested, while we look at the pure time to respond during the test phase in this additional hypothesis. When comparing the different studies, we found that the RTs in Experiment 4a were averagely 106 ms slower than studies finding heightened accessibility (Experiment 1-3), and approximately 255 ms slower than the studies displaying reduced accessibility (Higgins et al., 2020; Johnson et al., 2013; Langerock et al., 2021; see figure C). Therefore, in addition to our main hypothesis, we wanted to investigate whether the accessibility of an item in the focus of attention can change depending on the RT to this item, such that when RTs are very short the item would be less accessible, while when the RTs are a bit longer the item would be more accessible, and when the RTs are very long the accessibility of the item in the focus of attention is equal to another item in working memory. To investigate this, we presented more discriminable nonwords to make the lexical decision judgment easier for participants in an attempt to shorten RTs in Experiment 4b. The results of Experiment 4b demonstrate that we were able to decrease the mean RTs and thus, these mean RTs fall within the boundaries of other experiments finding heightened accessibility (see Fig. C), however, the analysis showed that the RTs for refreshed and unrefreshed probes were equal. Thus, it seems

that the accessibility of an item in the focus of attention does not depend on the length of the RT to the presented probe.

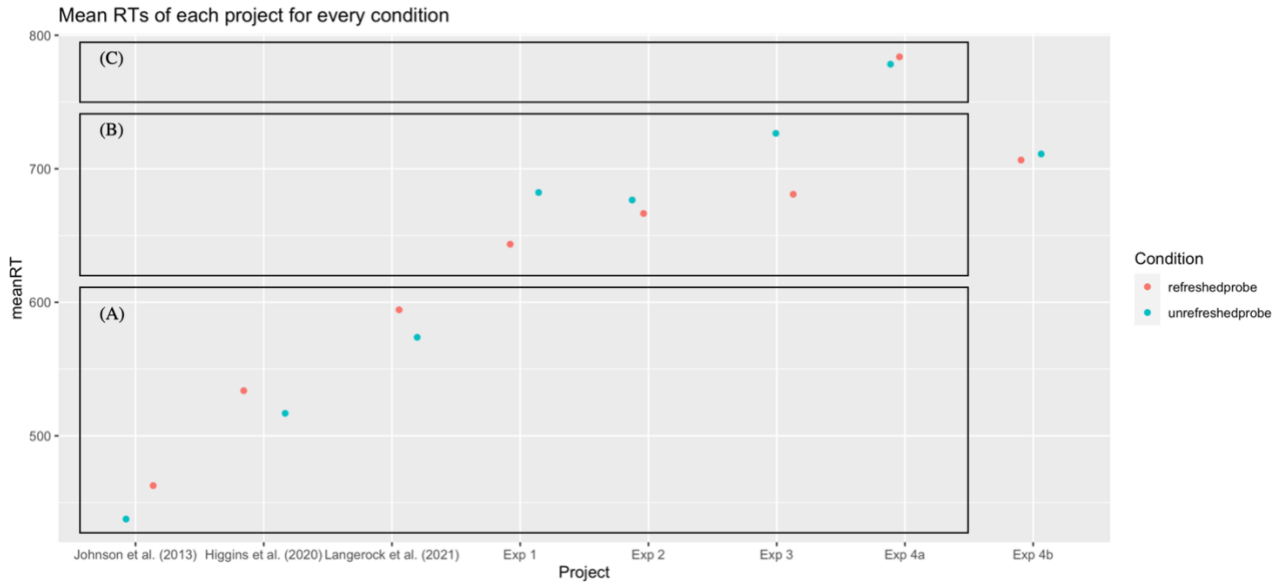

Figure C. Overall mean RTs fluctuations for the refreshed and unrefreshed probe for (A) experiments finding reduced accessibility for an item in the focus of attention compared to another item in working memory, (B) experiments finding heightened accessibility, and (C) experiments in which the RTs to the refreshed and unrefreshed probes were equal. On the right, the mean RTs of the refreshed and unrefreshed probes from Experiment 4b are presented which fall in the boundaries of experiments finding heightened accessibility (B), however, this experiment did not result in heightened accessibility, but in equal accessibility for the refreshed and unrefreshed probes (C).

Another potential explanation for the observed RT pattern (see Figure 6 in the main text) could be that in one task, participants strategically drop the refreshed item from the focus of attention (resulting in an inhibitory effect) while in another task, participants strategically keep the refreshed item in the focus of attention (resulting in a facilitative effect). One could assume that easier tasks (like an automatized reading as used by Johnson et al., 2013) would be associated with dropping the refreshed item from the focus of attention, thereby resulting in an inhibitory effect, whereas harder tasks (like a lexical decision as used in the current experiments) would be associated with keeping the refreshed item in the focus of attention, thereby resulting in a facilitative effect. However, as can be seen in Figure C, the pattern of results does not seem to vary with overall RT in a straightforward way. For example, in Experiment 4a, using a lexical decision task that resulted in rather slow responses, we did not

observe a facilitative effect (and descriptively, the effect was more similar to an inhibitory effect). While we do not want to claim that this is definitive evidence against the strategic account, we think that this pattern is not very consistent with it. Either way, our overall conclusion still holds that the inhibition-of-return-like effect is not as general as expected.

Supplementary material E

Additional Experiment

In this experiment, we aimed to investigate one final hypothesis to try to explain the inhibition-of-return-like effect found by Johnson et al. (2013), as this contrasted with the results of our Experiments 1-4b. More specifically, we speculated that the inhibition-of-return-like effect might have resulted from a combination of procedural aspects of the task (i.e., the response way, the to-be-executed task, and the representation of the word). To investigate this, we presented the original paradigm by Johnson et al. (2013) in a reversed way, such that all procedural aspects remained the same, but there was no longer any influence of the focus of attention. The experiment's design, hypothesis, and analysis plan were preregistered on OSF, see https://osf.io/h5zrs.

Method

Participants

We used Bayesian sequential hypothesis testing to determine the number of participants. We started with 30 participants and continued to increase by five participants (max. 60 participants) until a BF of 10 was reached for our main t-test, measuring for an inhibition-of-return-like effect for an item in the focus of attention. In total, 60 participants (25 female, 4 male, 1 did not respond, mean age = 24.37 years) from the University of Geneva took part in this experiment in exchange for course credits. The demographical information was collected for only 30 out of 60 participants, but like our previous experiments, all tested individuals are from the broad university community and thus, mostly young adults between 18 and 30 years old. Of these 60 participants, 17 were excluded because they did not reach the preregistered criterium of 75% valid trials (i.e., RTs > 150 ms and correctly responded to the probe and the refreshing cue). Of these 17 excluded participants, 3 did not reach 75% valid trials due to technical difficulties (less than 50% recorded responses). This resulted in a final sample of 43 participants.

Procedure & Materials

This experiment was programmed using E-prime 3.0 software (Psychology Software Tools, Pittsburgh, PA) and we used a Cedrus SV-1 voice key to record the onset time of speech production of the refreshed word. The materials remained the same as used in the replication study of Langerock et al. (2021), however, the procedure diverged from this replication study, and thus, the original paradigm by Johnson et al. (2013).

Participants were first presented with two words, one presented just above the center of the screen, the other just below the center of the screen during 1500 ms. They were asked to read these words silently. A blank screen followed for 500 ms. Up until here, the paradigm was the same as used in the study by Johnson et al. (2013), as well as in Experiment 1-4b. From here on, the procedure diverges. A probe was presented for 1500 ms, for which participants were instructed to read it aloud as fast as possible (i.e., task 1; previously presented at the end of the trial). Following a 100 ms blank screen, a refreshing cue was shown for 1500 ms pointing either to the top or bottom of the screen, indicating participants to refresh the word previously presented on the top or bottom of the screen (i.e., before probe presentation) and say it aloud as fast as possible (i.e., task 2; previously presented in the middle of the trial). Even though the probe and refreshing cue do not align with the meaning of these terms in a typical working memory task, we kept the same terms for the purpose of this experiment. The instructions remained the same as in our in-lab replication of the study by Johnson et al. (2013; Langerock et al., 2021) but were now reversed, following the current task order (i.e., probe presentation followed by refreshing cue).

The probe (i.e., task 1) could be (1) the to-be-cued word in task 2, (2) the not-to-be-cued word in task 2, or (3) a novel word. The novel word is a new word, that was not part of the memory items and that will not be cued in task 2. In this experiment, there are three kinds of trials. First, when the probe in task 1 is the same as the to-be-refreshed word in task 2, the

trial is a same read-refresh trial (i.e., read probe). Thus, the same word is read aloud as a probe in task 1 and is being refreshed in task 2. Second, when the probe in task 1 is the not-to-be-refreshed word in task 2 (meaning that the other memory item will be refreshed in task 2), the trial is a not-same read-refresh trial (i.e., unread probe). Thus, the word being read aloud as a probe in task 1 is not the same as the one that is being refreshed in task 2, however, both are memory items. Third, when the probe in task 1 is a completely new item, that was not part of the memory set, the trial is a novel read-refresh trial (i.e., novel probe). Thus, the word being read aloud as a probe in task 1 is a new word, not part of the memory set, while the to-be-refreshed item in task 2 is one of the two memory items. The novel read-refresh trials were not part of our main hypothesis but were included to keep the paradigm as similar as possible to the studies finding an inhibition-of-return-like effect in working memory (Higgins et al., 2020; Johnson et al., 2013; Langerock et al., 2021), as well as to be used in potential exploratory/control analyses.

To record RTs, a voice key was used to track response onset times to the refreshing cue (in task 2). To measure the accuracy for both the presented probe (task 1) and the refreshing cue (task 2), the experimenter remained next to the participant during the task to manually record whether the participant said aloud the correct probe (task 1) and said aloud the correct to-be-refreshed word (task 2). In total, there were 144 trials. There was an equal distribution of the three probe types (48 of each) and the trials were randomly displayed one after the other, with an intertrial interval of 3000 ms. Every 48 trials, there was a pause (so twice during the experiment). Before the start of the experimental trials, the task was explained to the participants, showing a concrete example. Then, the participant performed six training trials, two for each probe type, and one for each arrow direction (up or down).

The wordlist of 336 French words necessary for this experiment was the same as used in Langerock et al. (2021) and was created using the following procedure. Within the Lexique

3 database (New, 2006), a set of nouns was selected all consisting of six letters, two syllables, four or five phonemes, and a frequency (book frequency) between two and 50. This resulted in a list of 747 words. Plural nouns were taken out, as well as ambiguous words, words with an obvious negative connotation, words with double significations, words derived from English words, words pronounced exactly the same as another word in the word list, and derivations of verbs). This left us with 396 words, of which the 336 most frequent words were selected for the experiment. Of the remaining 60 words, the most frequent 16 words were used for the training.

Each participant received one of the fourteen different versions of the experiment. Seven word lists of 48 words were selected to serve as the novel probes in different versions of the experiment. These lists were created in such a way that the mean frequency was equal in the seven lists and equal to the remaining 288 words in the general wordlist. Each of these seven lists was used in two versions to serve as the "novel" probes. The remaining words in the wordlist (288 words) served as the to-be-remembered words and the refreshed and unrefreshed probes. Within the experiment, the order of presentation of the words remained the same (except for the 48 words taken out to serve as novel probes) in seven of the fourteen versions, and in the other seven versions of the experiment, only the order of the word pairs presented was switched, i.e., a word appearing on the first trial as the word above the center was now presented as the word below the center and vice versa. The order of the words was in function of their frequency, starting with the highest frequent words. This was the case in the general word list, as well as in the novel probe list.

The order of the different probe types (read probe, unread probe, or novel probe) was completely random, and the order of the pointing direction of the arrow was the same for all participants and based on a randomization defined beforehand (each arrow direction occurring 72 times during the experiment). This kind of design allowed the experimenter to verify directly

whether the participants read aloud the correct probe and to-be-refreshed word. The experimenter sat next to the participants during the experiment and registered 1) whether the participant read aloud the presented probe correctly in task 1, 2) whether the participant said aloud the correct to-be-refreshed word in task 2, 3) whether no supplementary noise was present during the trial (e.g., coughing or other noises that interfered with the correct detection of the response onset time for the refreshed word in task 2 by the voice key). The experiment script, raw data, and analysis for this experiment can be found on OSF (https://osf.io/e3x4d/).

Data analysis & Results

All trials in which the participant responds incorrectly or when no response was given or registered by the voice key within 1500 ms were excluded. Furthermore, following our preregistration, trials were excluded if noise was present (e.g., sneeze, cough, …) or if the RTs < 150 ms for the refreshing cue (i.e., task 2). Additionally, trials including a novel probe presented in task 1 were excluded from the dataset since these trials were only included to avoid any additional changes to the paradigm, relative to the original study of Johnson et al. (2013). Following these exclusion criteria, there was an exclusion of 10% of the trials; 9% read probe and 10% unread probe.

Following our preregistration, we ran a Bayesian paired one-sided t-test investigating an inhibition-of-return-like effect (i.e., RT read > RT unread). This resulted in moderate evidence (BF_{01} = 5.06) against an inhibition-of-return-like effect. As we did not find the latter effect, we continued with our preregistered, exploratory analysis. Firstly, we did a Bayesian paired one-sided t-test to investigate if there is a facilitative effect (i.e., RT read < RT unread). This showed that there is moderate evidence against this hypothesis (BF_{01} = 7.13). Secondly, we ran a Bayesian paired two-sided t-test to see whether there is any difference between the RTs to the read words and the unread words (RT read ≠ RT unread). The results of this t-test showed, again, moderate evidence against this hypothesis (BF_{01} = 5.92), meaning that the RTs

to the read words (mean RT = 768 ms, SE = 14 ms) are more or less equal to the RTs of the unread words (mean RT = 767, SE = 14 ms; see Fig. D). Thus, it seems that regardless of whether or not participants have read the word in task 1 before refreshing it in task 2, the RTs to this second task remained the same. Participants are not faster, nor slower in responding to a memory item that they just read before compared to a memory item that they had not just read before.

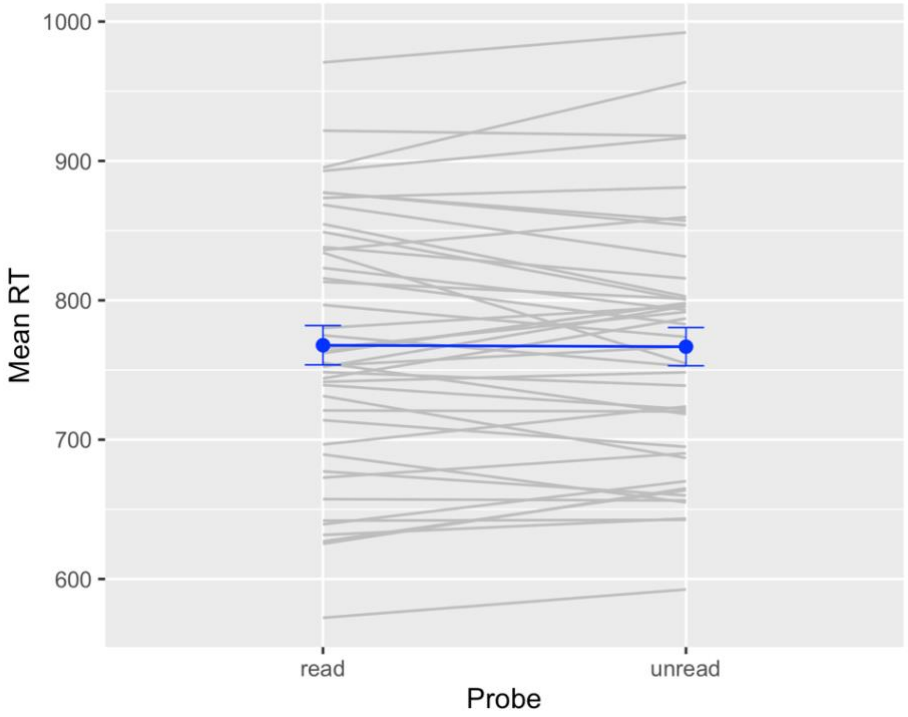

Figure D. Mean RT (in ms) to the read and unread probes in Experiment 5 (in blue), presented with error bars (SE) and individual mean RT (in grey).

In conclusion, when repeating the task by Johnson et al. (2013) but inverting the order of the task, we did not obtain the same results. More specifically, in the task by Johnson et al. (2013), participants were instructed to memorize two words after which they were instructed to refresh one of these words and finally, to read aloud the presented word. When this presented word was the same as the refreshed word, participants were slower to say this word aloud compared to when this word was the other memory item (i.e., unrefreshed word). In the current experiment, participants were also instructed to memorize two words after which they were

asked to read aloud the presented word, followed by the instruction to refresh one of the two previously presented words (i.e., a reversed version of the task by Johnson et al., 2013). In this reversed experiment, participants were equally fast to refresh a word that they had previously read as to refresh a word that they had not previously read. Based on these results, it seems that the inhibition-of-return-like effect does not result from a combination of procedural aspects of the task (i.e., the response way, the to-be-executed task, and the representation of the word).