The influence of working memory mechanisms on false memories in immediate and delayed tests

Marlène Abadie, Christelle Guette, Amélie Troubat, Valérie Camos

Cognition 252 (2024) 105901

Contents lists available at ScienceDirect

Cognition

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

The influence of working memory mechanisms on false memories in immediate and delayed tests

Marlène Abadie [a,*], Christelle Guette [a], Amélie Troubat [a], Valérie Camos [b]

[a] Aix Marseille Université, CNRS, CRPN, Marseille, France
[b] Département de Psychologie, Université de Fribourg, Switzerland

ARTICLE INFO

Keywords:
False memory
Working memory
Long-term memory
Recall
Immediate test
Delayed test

ABSTRACT

There is growing evidence that false memories can occur in working memory (WM) tasks with only a few semantically related words and seconds between study and test. Abadie and Camos (2019) proposed a new model to explain the formation of false memories by describing the role of articulatory rehearsal and attentional refreshing, the two main mechanisms for actively maintaining information in WM. However, this model has only been tested in recognition tasks. In the present study, we report four experiments testing the model in recall tasks in which the active maintenance of information in WM plays a more important role for retrieval. Short lists of semantically related items were held for a short retention interval filled with a concurrent task that either impaired or not the use of each of the WM maintenance mechanisms. Participants were asked to recall the items immediately after the concurrent task (immediate test) or later, at the end of a block of several trials (delayed test). In the immediate test, semantic errors were more frequent when WM maintenance was impaired. Specifically, rehearsal prevented the occurrence of semantic errors in the immediate test, while refreshing had no effect on their occurrence in this test, but increased semantic errors produced only in the delayed test. These results support Abadie and Camos (2019) model and go further by demonstrating the role of active information maintenance in WM in the emergence of false memories. The implications of these findings for understanding WM-LTM relationships are discussed.

1. Introduction

"All remembering is constructive in nature," Roediger and McDermott (1995) pointed out in their most cited paper introducing the technique, referred to as the Deese, Roediger and McDermott (DRM) paradigm, that would become the most popular for the experimental investigation of false memories. The basic procedure is straightforward: Participants hear or read lists of 12 to 15 words that are the strongest semantic associates of an unpresented critical word according to word association norms (e.g., Nelson et al., 2004). These words such as "hill, valley, climb, summit, top, etc." share many semantic associations with each other and with a critical theme word not presented, "mountain" (Brainerd et al., 2008). When tested immediately after the presentation of each list or of several lists, participants often recall and/or recognize the theme word as having been presented, resulting in a robust form of memory illusion. The majority of studies have used long lists of semantically related items and relatively long delays between the study of the material and its later testing (see Gallo, 2006; Gallo, 2010; Chang & Brainerd, 2021, for reviews). This phenomenon has been replicated across a variety of experimental situations and is now considered a hallmark of long-term memory (LTM) functioning.

Recent studies have extended these settled findings by showing that the DRM illusion can also occur in working memory (WM) tasks for lists of only a few items when a short interval (e.g., 4 s) is provided between study and test (e.g., Atkins & Reuter-Lorenz, 2008). At first glance, this striking parallel between the memory distortions that can affect both short- and long-term tests suggests, as Roediger and McDermott (1995) had hinted, that reconstructive retrieval processes are not exclusive to LTM, but can also operate in WM. WM is conceived as a limited capacity system for maintaining relevant representations in the short term. This apparent convergence between WM and LTM processes supports a unitary view of memory. However, most of these studies used recognition tasks that are questionable when trying to distinguish the influence of WM processes from that of LTM processes because they favor the use of LTM (e.g., Uittenhove et al., 2019). Furthermore, although still using recognition tasks, several experiments have shown that short-term false

* Corresponding author at: Centre de Recherche en Psychologie et Neurosciences, Bâtiment 9 Case D, 3 place Victor Hugo, 13331 Marseille Cedex 3, France.
E-mail addresses: marlene.abadie@univ-amu.fr (M. Abadie), christelle.guette@univ-amu.fr (C. Guette), valerie.camos@unifr.ch (V. Camos).

https://doi.org/10.1016/j.cognition.2024.105901
Received 25 October 2023; Received in revised form 17 May 2024; Accepted 19 July 2024
Available online 15 August 2024

memories only emerge when the opportunities to maintain information in WM are reduced (e.g., Abadie & Camos, 2019). These findings raise two questions of fundamental importance, the first concerning the involvement of WM in short- and long-term false memories, and the second concerning the similarities and dissociations between WM and LTM processes.

The present study was designed to address these two questions using two novel approaches. The first is to shift from recognition to recall. Compared to recognition, recall is more conducive to the use of active maintenance in WM and of recollective processes during retrieval (e.g., Uittenhove et al., 2019). These are optimal conditions for testing whether short-term false memories actually arise during WM maintenance. The second approach, inspired by Abadie and Camos (2019) methodology, involves orthogonally varying the availability of the two main WM maintenance mechanisms, articulatory rehearsal and attentional refreshing, to compare the occurrence of short- and long-term false memories when WM maintenance via one and/or the other of these mechanisms is impeded and when it is not. Abadie and Camos (2019) proposed that the use of each of these WM maintenance mechanisms moderates the occurrence of short- and long-term false memories differently. Articulatory rehearsal would reduce short-term false memories without affecting long-term false memories, while attentional refreshing would affect only long-term false memories. The four reported experiments showed exactly this pattern. Before presenting the experiments, we review previous evidence of short-term false memories in recognition, present Abadie and Camos' theoretical proposal that accounts for them, describe the pitfalls of previous experimental tests and sketch the extension of Abadie and Camos model to encompass false memories in recall.

1.1. False memories in WM

Over the past decades, several studies, most of which using recognition tasks, have demonstrated that false memories may occur rapidly, on the time scale of a WM task, and may not require the use of long lists of memory items. Coane et al. (2007) showed that after studying DRM lists with only five to seven items and a delay of about 1 s, participants incorrectly endorsed the unpresented critical theme word as "old" 20–22% of the time. Although false alarm rates for critical lures were lower than those observed in long-term studies (e.g., 40–55%; Roediger & McDermott, 1995), they were significantly higher than false alarm rates for other weakly related or unrelated lures. Moreover, a semantic interference effect was found on reaction times (RTs) for correct rejections of critical lures. Compared with unrelated distractors, critical lures took approximately 100 ms or more to be rejected, whatever the sizes (5–7 items) of the memory lists. Atkins and Reuter-Lorenz (2008, Exp. 1A & 2A) also reported fairly high false alarm rates (31%) for critical lures and increased RTs for their correct rejection in a short-term recognition test[1] with 4-item DRM lists that was completed after a 3–4 s retention interval. These findings suggest that false recognition errors can reliably occur within the temporal (i.e., after a few seconds of retention) and set size (i.e., for short memory lists) parameters characteristic of WM.

A subsequent study showed that the same mechanisms involved in long-term false memories underlie the semantic interference effect in a short-term recognition test (Atkins & Reuter-Lorenz, 2011). Further studies (Flegal et al., 2010) compared false recognition rates, confidence ratings and Remember/Know judgments (Roediger III et al., 1993) in an immediate recognition test administered after a 3–4 s retention interval and in a surprise delayed recognition test that occurred approximately 20 min later. Critically, the incidence of false recognition of semantically related lures was found to be relatively stable across delays (13–23%). Moreover, neither confidence ratings nor "remember" judgments associated with false recognition changed over time, suggesting that the subjective feeling of certainty that characterizes some cases of false recognition may be relatively invariant over time.

Although these studies suggest some overlap in the mechanisms of LTM and WM, other studies have shown dissociations between immediate and delayed tests on false memory rates. Flegal and Reuter-Lorenz (2014) conducted two experiments in which the level of information processing (Craik & Lockhart, 1972) was manipulated as participants encoded 4-item DRM lists. Level of processing had little effect on recognition performance at delays of a few seconds. However, deep encoding (relative to shallow encoding) increased true and false memory rates in a surprise recognition test 20 min later. Furthermore, some studies have varied the presence or absence of a concurrent task during the retention interval between the study phase and the immediate test. For example, in Atkins and Reuter-Lorenz's (2008, Exp.1 A & 2 A) study, this interval was either filled or not filled by a mathematical distraction task requiring attention. Interestingly, although false recognition still occurred, it was greatly reduced in the absence of a concurrent task during the retention interval. These latter findings question the role of WM in the occurrence of false memories. The emergence of short-term false memories may be attributed to the fact that performing a concurrent task during the retention interval impairs the maintenance of information in WM. Hence, providing compelling evidence that WM produces memory distortions requires first ensuring that items can be maintained in and retrieved from WM.

1.2. WM maintenance mechanisms and their effects in immediate and delayed tests

The main point of divergence between the different WM models is the question of the separability of LTM and WM (see Logie et al., 2021, for a review). While some propose that WM is the activated part of LTM (Cowan, 1999; Engle, 2001; Oberauer, 2002), others envision a clear distinction between the two memory systems (Baddeley, 1986, 2007; Barrouillet & Camos, 2015; Unsworth & Engle, 2007). In the latter view, WM would construct, maintain, and transform mental representations according to the goals of the task at hand. Constructions and transformations would often require the use of information stored in LTM and could also lead to the creation and storage of new information in LTM.

Within the latter conception, Barrouillet and Camos (2015) time-based resource-sharing (TBRS) model of WM describes two main mechanisms for maintaining representations in WM: articulatory rehearsal and attentional refreshing. Although the two mechanisms appear similar in many ways, they differ in some important ways (Camos, 2015, 2017; Camos et al., 2018). First, rehearsal is assumed to rely on subvocal articulation of verbal information, whereas refreshing is assumed to rely on attentional reactivation of memory traces. Second, each of these mechanisms is assumed to maintain distinct traces of items in separable subsystems of WM. On the one hand, rehearsal is seen, as in Baddeley (1986); Baddeley (2007) model, as a speech-based maintenance mechanism that can only be used to maintain verbal information. It operates through articulatory repetition of the phonological traces of verbal items that are temporarily stored in a phonological loop. On the other hand, refreshing is thought to reactivate memory traces of any type (verbal or nonverbal) in an executive loop by paying attention to them. Finally, in contrast to refreshing, which is an attention-based process, articulatory rehearsal is thought not to rely, or to rely only minimally, on attention (e.g., Camos & Barrouillet, 2014; Chen & Cowan, 2009; Naveh-Benjamin & Jonides, 1984).

Numerous behavioral, developmental and neuroimaging studies have shown that attentional refreshing and articulatory rehearsal are

[1] Note that we use the terms short-term test or immediate test interchangeably, as is the case in the WM literature, to refer to memory tests that take place after a few seconds of retention, which may or may not be filled with a concurrent task. Conversely, we use the terms long-term test or delayed test to refer to memory tests that take place after several trials with an immediate memory test.

two independent processes. For example, Camos et al. (2009, 2011) showed that varying the opportunity to use one mechanism while controlling the other during the retention interval of a typical WM task resulted in poorer immediate recall performance. Moreover, orthogonal manipulation of the two mechanisms produced an additive effect on correct immediate recall, suggesting that both mechanisms contribute to recall performance, but independently. Further supporting the independence of the two mechanisms, brain imaging studies have shown that different brain networks underlie each of the two mechanisms (e.g., Johnson et al., 2005; Raye et al., 2007; Smith & Jonides, 1999; Trost & Gruber, 2012). Additionally, they have different effects on long-term maintenance. While rehearsal has no effect on delayed recall (Greene, 1987), increasing the opportunity to use refreshing improves delayed recall (Camos & Portrat, 2015; Loaiza & McCabe, 2012, 2013). To account for these different effects of maintenance mechanisms on delayed recall, the TBRS model proposes that rehearsal restores sensory input through output planning processes (Jones et al., 1995, 2007; Macken et al., 2016), which does not rely on LTM traces after the initial configuration of the articulatory program. Refreshing, on the other hand, would leave partial traces in LTM of reconstructions it made in WM. In other words, the use of rehearsal does not affect LTM, whereas refreshing favors the long-term maintenance of memory traces previously maintained in WM.

1.3. A new account of short and long-term false memories

To examine the role of WM maintenance mechanisms on false memory, Abadie and Camos (2019) conducted a series of experiments in which the availability of each mechanism was manipulated. The paradigm was similar to that used in the previously presented short-term false memory studies (Atkins et al., 2008; Flegal et al., 2010), except that the attentional demand of the concurrent task performed during the retention interval was varied in order to manipulate refreshing opportunities. The concurrent task was either high (e.g., verifying mathematical operations) or low (e.g., simply reading of the same operations) attentional demanding. Increasing the attentional demand of a concurrent task while maintaining items reduces the opportunities for refreshing the memory items (e.g., Barrouillet et al., 2004, 2007). In addition, to vary the availability of rehearsal, the concurrent task was performed aloud (i.e., articulatory suppression) or silently. Performing a memory task under articulatory suppression reduces the opportunity to rehearse the memory items (e.g., Baddeley, 1986, 2012). Results from an immediate recognition test administered after the 4-s retention interval showed that false recognition of semantically related items was higher when the use of both WM maintenance mechanisms was impeded than when it was not. In contrast, in a delayed test administered later at the end of the experiment, false memories were less frequent when both WM mechanisms were impeded than when they were not. These findings show a dissociation between immediate and delayed tests on false memory rates. They suggest that false memories found in immediate tests do not arise from WM, because these errors were less frequent when items were actively maintained in WM. This series of experiments also unraveled the role of each mechanism. Articulatory suppression increased false memories in the immediate test, but had no effect in the delayed test. In contrast, reducing refreshing opportunities had no effect on false memories in the immediate test, but reduced false memories in the delayed test. Although we are not aware of other studies that have examined the effect of refreshing, these findings are consistent with some studies that also report a larger false memory effect under articulatory suppression (Atkins et al., 2011; Macé & Caza, 2011).

To account for the emergence of short- and long-term false memories, Abadie and Camos (2019) proposed a model that integrates the fuzzy-trace theory (FTT, Reyna & Brainerd, 1995) conception of LTM with the TBRS (Barrouillet & Camos, 2015) model of WM. According to the FTT, the nature of representations retrieved during a memory test, either verbatim or gist, determines the formation of false memories. Verbatim traces are representations of the surface forms of the items to be remembered (e.g., the details that accompany their presentation), whereas gist traces are representations of the semantic features of the items (e.g., Brainerd et al., 2014). Both support true memory for experienced items, but their retrieval has opposite effects on false memory. In recognition tasks, when related distractors are presented at test, retrieval of verbatim memories leads to their rejection, whereas retrieval of gist memories leads to their acceptance. Therefore, variations in the contribution of these two types of traces impact the incidence of false memories. By integrating the TBRS model with the FTT, Abadie and Camos (2019) model depicts how WM maintenance mechanisms influence the formation of verbatim and gist traces and thus the emergence of short- and long-term false memories.

Based on the literature on the two WM maintenance mechanisms reviewed above, Abadie and Camos (2019) model, summarized in Fig. 1, proposes that the use of articulatory rehearsal strongly emphasizes verbatim memory and minimizes reliance on gist memory. Rehearsal does not rely on LTM and does not use semantic information from the items (see also Loaiza & Camos, 2018). Thus, the use of rehearsal should reduce the occurrence of short-term false memories and increase true memories on immediate tests but should have no effect on LTM and thus on delayed tests. In contrast, refreshing should allow for the maintenance of verbatim and gist traces in WM, and its disruption should reduce verbatim and gist retrieval. Because gist traces are countered in the short term by strong verbatim traces, the use of refreshing should increase correct recognition but have no effect on false memories in immediate tests. However, its use should increase both true and false memory in delayed tests. The reconstructions made by refreshing traces in WM leave partial traces in LTM, thus favoring their correct retrieval in delayed tests. Moreover, gist traces are less susceptible to loss and interference than verbatim traces (e.g., Abadie et al., 2013, 2017). They more strongly support memory retrieval after prolonged retention, thus favoring false memories. The predictions of this integrated model were supported by three series of experiments that examined recognition performance in immediate and delayed tests and provided direct measures of verbatim and gist representations in adults and children (Abadie & Camos, 2019; Abadie & Rousselle, 2023; Rousselle et al., 2022).

1.4. The case of recall

Most previous studies have examined short-term false memories using recognition tasks. However, it is now essential to examine whether these false memories also occur in recall tasks for at least three reasons. First, in the traditional DRM paradigm, false memories also manifest as semantic errors in free recall. Therefore, it is appropriate to examine whether short-term false memories also occur in recall tasks. This is particularly important because recognition is more sensitive to the DRM illusion than recall (e.g., Chang & Brainerd, 2021). In two experiments, Atkins and Reuter-Lorenz (2008; Exp. 1B & 2B) demonstrated the occurrence of semantic errors in a short-term recall task. Although small in number (about 5% compared to about 20% in recognition), these errors were more frequent than other types of errors. This suggests that, as in long-term studies, the short-term false memory effect is stronger in recognition than in recall. Thus, the almost exclusive use of recognition in previous studies may have led to an overestimation of the occurrence of short-term false memories.

Second, it is important to ensure that short-term false memories are not strictly dependent on specific retrieval processes. If these errors do not occur in recall, this would imply that they are inherent to the retrieval processes specifically involved in recognition. Whether recall and recognition rely on different retrieval processes is still a matter of debate in the LTM literature. However, a number of studies suggest that both recollective processes, which involve direct access to verbatim traces of the items studied, and non-recollective processes, which involve a reconstruction operation accompanied by a familiarity judgment, would be involved in recall. Conversely, recognition would rely

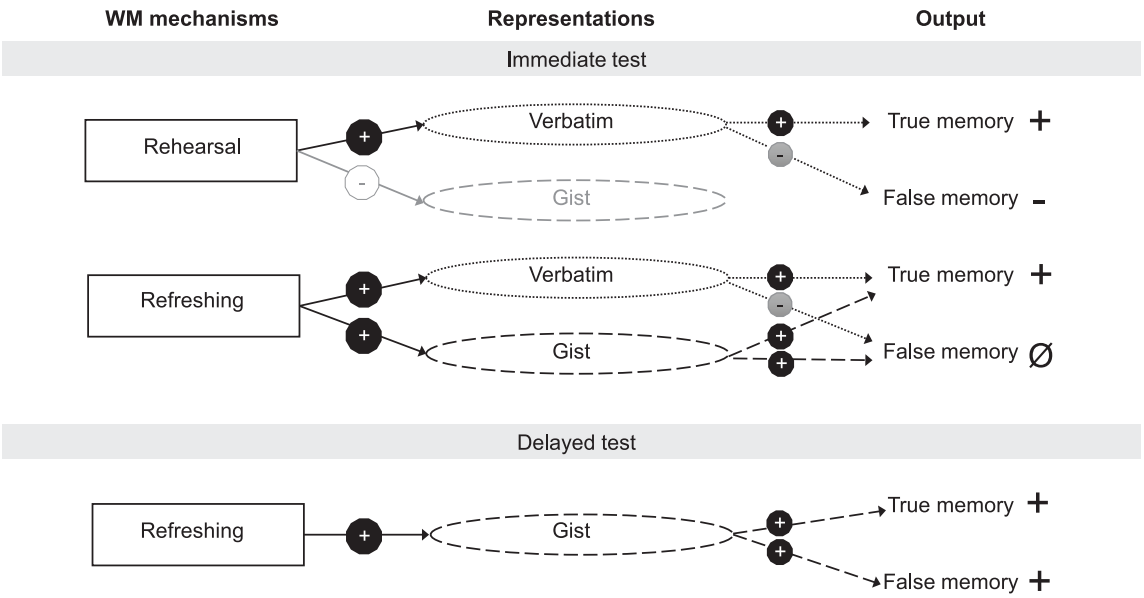

Fig. 1. *Illustration of the model proposed by* Abadie and Camos (2019).
Note. According to Abadie and Camos (2019), the use of rehearsal promotes short-term maintenance of verbatim memory and decreases reliance on gist memory. Verbatim retrieval increases true memory and decreases false memory in the immediate test. The use of refreshing promotes short-term maintenance of verbatim and gist memory. In the immediate test, the retrieval of strong verbatim traces increases true memory and counteracts the effect of gist memory on false memory. Refreshing also promotes the retrieval long-term gist traces, increasing both true and false memory.

more on a single non-recollective process (Brainerd et al., 2009; Brainerd et al., 2012; Brainerd et al., 2014; Brainerd et al., 2015; Brainerd & Reyna, 2010; Malmberg, 2008; Wixted, 2007). Thus, short-term false memories may have been overestimated in recognition tests favoring the use of non-recollective processes.

Finally, recent studies in the WM literature have suggested that recognition requires active maintenance of the memory items to a lesser extent than recall (Allen et al., 2018; Uittenhove et al., 2019). If recognition tests only weakly require active maintenance in WM, the conclusions of studies on short-term false memories that have primarily used recognition tests may need to be reconsidered. Moreover, to test the respective contribution of the mechanisms responsible for active maintenance in WM, it is necessary to ensure that active maintenance is strongly involved in the task used. Thus, the use of recall tasks is an essential condition to draw sound conclusions about false memories, to assess the role of WM in their occurrence, and to reevaluate the proposal of Abadie and Camos (2019).

1.5. The present study

We reported four experiments in which participants had to maintain four-item DRM lists over a retention interval of a few seconds. In Experiments 1 and 2, the availability of articulatory rehearsal and of attentional refreshing during the retention interval was manipulated orthogonally. Then, participants had to recall the four words immediately after the retention interval and in a delayed test at the end of the experiment. Experiments 3 and 4 focused more specifically on refreshing, introducing stronger manipulations of the mechanism. Moreover, a procedure to assess participants' confidence level or subjective experiences associated with each recall was introduced into these experiments.

As predicted by Abadie and Camos (2019) model, because rehearsal strengthens verbatim traces in WM, its use should facilitate direct access to these traces, increasing true recall and reducing the incidence of semantic false recall in the immediate test. In addition, as rehearsal has only short-term and not long-term effects, it should not impact true and false delayed recall. In contrast to rehearsal, refreshing allows information to be maintained not only at short, but also at long delays. It also enhances gist retrieval in delayed recognition tests. Therefore, we expected that the use of refreshing would increase true recall in the immediate test with no impact on the occurrence of semantic errors. In the delayed test, the use of refreshing should foster gist retrieval and thus the occurrence of semantic errors. The predictions of the four experiments and the main results obtained are summarized in Table 1.

2. Experiment 1

2.1. Method

2.1.1. Ethics and sample size

All the experiments were carried out in accordance with the recommendations of the Declaration of Helsinki. Study procedures were approved by the Aix-Marseille University Institutional Review Board, and all participants gave written informed consent prior to participation.

Based on the effect size of varying articulatory rehearsal opportunities on false memories obtained on recognition tests (Cohen's $d = 1.18$) in a previous study using a similar method (Abadie & Camos, 2019), a power analysis indicated that a total sample size of 40 participants would be required to achieve a 95% power (G*Power, Faul et al., 2007). We collected 40 or more participants per experiment to account for potential data loss. Nevertheless, we performed Bayesian analyses in which evidence of a null effect is equally informative (Kruschke, 2011; Kruschke & Liddell, 2018), and Type I error does not increase with optional stopping (Rouder, 2014).

2.1.2. Participants

Forty young adults ($M_{age} = 21.63$ years, $SD_{age} = 2.46$, range = 19–30 years old; 37 females, 3 males) took part in Experiment 1. They were students at Aix-Marseille University and received course credits in return for their participation. All participants declared that they were healthy, had no learning disabilities and were native French speakers.

2.1.3. Materials

2.1.3.1. Word lists. We selected 40 semantically related word lists from

Table 1
Summary of predictions and main results for recall performance in the immediate and delayed tests in the four experiments.

Predictions	Experiment 1		Experiment 2		Experiment 3		Experiment 4	
	Immediate test	Delayed test	Immediate test	Delayed test	Immediate test	Delayed test	Immediate test	Delayed test
Correct recall								
Rehearsal increases correct recall in the immediate test only	✓	✓ (✓)	✓	✓ (✓)				
Refreshing increases correct recall in the immediate and delayed tests	✓	× (×)	✓	× (×)	✓	× (×)	✓	✓ (×)
Semantic errors								
Rehearsal decreases semantic errors in the immediate test only	✓	✓ (✓)	✓	✓ (✓)				
Refreshing increases semantic errors in the delayed test only	✓	× (✓)	✓	× (✓)	✓	× (✓)	✓	× (✓)

Note. ✓ indicates that there was substantial evidence for the prediction (i.e., a Bayes Factor > 3), ✓ indicates that there was weak evidence for the prediction (i.e., a Bayes Factor between 1 and 3), × indicates that there was substantial evidence against the prediction (i.e., a Bayes Factor < 0.33), and × indicates that there was weak evidence against the prediction (i.e., a Bayes Factor between 1 and 0.33). Results on the conditionalized delayed recall score (see the results section for more details on this score) are given in parentheses.

those used in Abadie and Camos (2019). Each list included the four words most strongly associated (e.g., "snow, ski, winter, downhill") with a given non-presented theme word (e.g., "sled"). Each list was selected on the basis that these four words had no cross-association with any other given list or their theme words. The 40 lists were then separated into four groups of 10 lists that were equated in mean Backward Associative Strength (BAS, $M = 0.35$ for each group of lists). Each group was assigned in a counterbalanced manner to each of the four experimental conditions (see the description of each condition below). The order of presentation of the lists within a given group, and for each participant, was randomized.

2.1.3.2. Concurrent task. Based on the study by Atkins and Reuter-Lorenz (2008), we used an operation verification task as a concurrent task, i.e., verifying whether a mathematical expression was true or false. Each mathematical expression consisted of two successive operations with first either a division or multiplication, and second an addition or subtraction, followed by a result to verify. Half of the mathematical expressions were shown with a true result (e.g., $4 \times 5 - 2 = 18$), the other half with a false one (e.g., $6 / 2 + 4 = 8$). There were 40 mathematical expressions in total, assigned in a counterbalanced manner to each of the four conditions, 10 per condition.

2.1.4. Procedure

Experiment 1 was presented on a laboratory computer using the *E*-Prime 3.0 software (Psychology Software Tools, Pittsburgh, PA). Before starting the experiment, participants were informed that this was a memory experiment and that they would be taking several recall tests.

As in previous studies (e.g., Abadie & Camos, 2019; Atkins & Reuter-Lorenz, 2008), at the beginning of each trial, four-word associates of a given list appeared simultaneously on the screen during 1750 ms (Fig. 2). Participants were asked to read the words aloud to ensure that they had enough time to read them. Then, participants had to complete the concurrent task while maintaining the four words for subsequent recall. A mathematical expression appeared on the screen for 4000 ms. Depending on the experimental condition, participants had to either verify the expression (true or false) by pressing the appropriate key (high attention demanding condition) or simply read it and press the space bar without verifying it (low demanding condition). Verifying whether the result of a mathematical expression is true or false requires more attention than simply pressing the space bar. The opportunities for refreshing the words were therefore lower in the first condition than in the second. Similar manipulations of refreshing opportunities have been shown to be effective in preventing the attentional maintenance of information in WM (e.g., Camos et al., 2018, for a review). Moreover, participants were either asked to perform the concurrent task aloud to reduce the use of articulatory rehearsal, or silently. Numerous studies have shown that concurrent articulation when maintaining verbal information strongly impedes the use of rehearsal to maintain it (e.g., Baddeley, 1986, 2012; for reviews). The availability of each WM maintenance mechanism, rehearsal and refreshing, was orthogonally manipulated, resulting in four different experimental conditions. Participants were then instructed to orally recall the four presented words (immediate recall test). There were 10 trials per condition. At the end of each condition, participants were asked to count backwards by twos from a randomly chosen number between 100 and 1000 for 1 min. They were then instructed to orally recall as many words as they could from the 40 words they had just seen in the last condition (delayed recall test). There were no time limits on the recall phases. Each participant completed the four conditions, the order of which being counterbalanced across participants.

Prior to the study phase, all participants underwent a training phase that included practice for the immediate and delayed tests. The experiment did not begin until the experimenter had ensured that all participants had understood the instructions. Finally, demographic information concerning the participants' gender and date of birth was collected orally by the experimenter at the end of the experiment.

2.2. Results and discussion

To ensure that participants complied with the instructions and paid sufficient attention to the concurrent tasks, the experimenter checked that they responded to all task trials. Participants who responded to <90% of the concurrent task trials and those for whom immediate recall performance differed from the average performance by more than 3SDs were excluded. This resulted in the exclusion of the data of four participants.

Bayesian analyses were conducted using JASP Version 0.16.3 (JASP Team, 2022). In Bayesian hypothesis testing, the strength of evidence for a specified model (M1) was quantified by comparing this model against a null or reduced model (M0). The ratio of the likelihood of the two models under comparison is the Bayes Factor (BF_{10}). BF_{10} of each model was obtained by comparing it to the null model. Strength of evidence is evaluated using Kass and Raftery (1995) interpretation of Bayes Factors.

2.2.1. Performance on the mathematical verification task

Performance was very good for the two conditions in which participants had to verify mathematical expressions. A Bayesian paired sample *t*-test provided substantial evidence against a difference as a function of whether the expressions were read aloud (83.9%, $SD = 14.6$) or silently

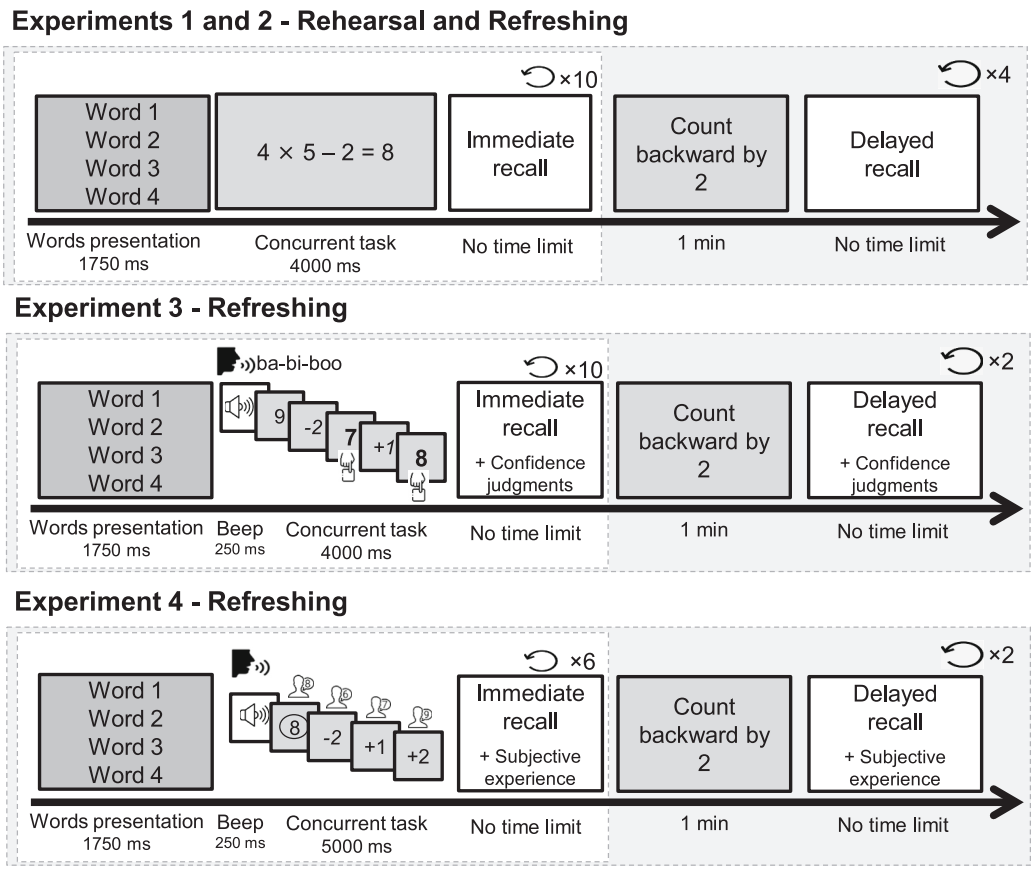

Fig. 2. Illustration of experimental procedure.
Note. Experiments 1 and 2 were designed to investigate the effects of rehearsal and refreshing on correct and incorrect recall, while Experiments 3 and 4 focused specifically on refreshing. In Experiment 1, depending on the attentional demand condition, participants' concurrent task was either to read and verify whether a mathematical expression was correct, or to read it and press the space bar. This task was performed either aloud (articulatory suppression) or in silence. The procedure in Experiment 2 was similar to that in Experiment 1, except that participants had to recall 40 words in the delayed recall task. In Experiment 3, depending on the attentional demand condition, participants' concurrent task consisted of solving several simple addition and subtraction problems or simply pressing a key each time a result appeared on the screen. The use of articulatory rehearsal was prevented in both conditions by having participants continuously utter the syllables "ba-bi-boo" aloud during the retention interval. In Experiment 4, depending on the attentional demand condition, the pace of the concurrent task was either high, with five operations to be processed in 5 s, or low, with only three operations. Articulatory rehearsal was blocked in both conditions by asking participants to perform the concurrent task aloud.

(81.4%, $SD = 15.9$; $BF_{10} = 0.22$).

2.2.2. Recall performance in the immediate test

2.2.2.1. Correct recall. To test our hypotheses about the effects of refreshing and rehearsal on correct recall in the immediate test (Table 1), a default Bayesian repeated-measures ANOVA was conducted on the percentage of correct recall with attentional demand of the concurrent task (high vs low) and the presence or not of articulatory suppression (read the mathematical expressions aloud vs. silently) as within-subject factors. We reported the best model, the one with the largest BF_{10}, and then decomposed the effects included in that model that were relevant to the hypotheses. Table 2 shows the percentage of correct recall as a function of experimental conditions.

The additive model including the main effects of attentional demand and articulatory suppression was the best ($BF_{10} = 2.23 \times 10^{23}$). As expected, there was decisive evidence that recall performance was better when participants performed the low demanding (88.3%, $SD = 8.58$) than the high demanding (80.6%, $SD = 11.4$) concurrent task and when they performed the concurrent task silently (90.1%, $SD = 8.44$) rather than aloud (78.8%, $SD = 9.89$). These results are consistent with those of many studies showing that orthogonal manipulation of rehearsal and refreshing yields additive effects. This suggests that both mechanisms contribute to recall performance in short-term tests, but do so independently (e.g., Camos, 2015; Camos et al., 2009, for reviews).

2.2.2.2. Recall errors. Incorrect responses were classified by two independent and trained raters into four types: semantic, phonological, intrusions and other errors. Interrater agreement was 72.5% before discussion among raters and full interrater agreement was achieved after discussion. When the word recalled was the theme word for the trial, an associate of that theme (i.e., not presented but listed on the original lists from which our four-items list was created; (Abadie and Camos, 2019) or a word related in meaning to two or more items presented in the list, it was classified as a semantic error.[2] When the recalled word differed from only one syllable or swapped two syllables from a studied item and was not related in meaning with an item in the list, it was classified as a phonological error. Errors were defined as intrusions when the recalled word was studied in a previous list. Finally, the category "other errors"

[2] The percentages of semantic errors by type (i.e., incorrect recall of the theme word or incorrect recall of an associate of that theme) are reported in the supplementary material in the OSF. In all experiments, the pattern of results remained the same whether all semantic errors or only incorrect recall of the theme word was considered. Therefore, the reported analyses were performed on all semantic errors.

Table 2
Mean percentage of correct recall and recall errors (semantic, phonological, intrusion and other errors) as a function of attentional demand of the concurrent task (high vs. low) and the presence or not of articulatory suppression (concurrent task performed aloud vs. silently) for the immediate and the delayed test in Experiments 1 and 2.

		Concurrent task			
		High demanding		Low demanding	
Experiment	Recall accuracy	Aloud	Silently	Aloud	Silently
		Immediate test			
	Correct recall	74.2 (9.1)	87 (9.8)	83.5 (8.5)	93.2 (5.4)
	Recall errors				
	Semantic	15.8 (7.1)	6.6 (6.4)	10.5 (7)	4.2 (3.4)
	Phonological	1.3 (1.8)	2.0 (2.3)	1.7 (2.2)	0.8 (1.9)
	Intrusion	5.1 (6.3)	3.1 (4.8)	1.6 (2.9)	1.0 (2)
	Other	3.5 (5)	1.3 (2.1)	2.8 (3.3)	0.8 (1.8)
Exp. 1		*Delayed test*			
	Correct recall	33.3 (10.3)	31.4 (14.4)	32.3 (11.8)	30.7 (13.5)
	Cond. correct recall	42.5 (10.5)	35.5 (15.2)	37.7 (13.8)	32.3 (14)
	Recall errors				
	Semantic	3.9 (2.9)	6.7 (5.2)	4.9 (4.3)	5.4 (4.1)
	Phonological	0.4 (1.1)	0.3 (0.9)	0.6 (1.1)	0.3 (0.8)
	Intrusion	2.2 (5.2)	1.2 (3.6)	1.7 (4.4)	0.8 (2.2)
	Other	0.8 (1.3)	0.8 (1.7)	0.9 (2.4)	1.5 (3.5)
		Immediate test			
	Correct recall	75.6 (7.8)	87.4 (8.6)	81.3 (7.7)	91.9 (7.1)
	Recall errors				
	Semantic	16 (6.6)	7.6 (5.2)	13.2 (7)	4.7 (4.9)
	Phonological	1.2 (1.3)	1.2 (1.7)	1.5 (1.7)	0.7 (1.4)
	Intrusion	4.0 (5.6)	1.9 (3.3)	1.8 (2.4)	1.3 (2.6)
	Other	3.2 (3.4)	1.9 (4.7)	2.2 (4.2)	1.5 (2.3)
Exp. 2		*Delayed test*			
	Correct recall	43.9 (11.1)	43.5 (14.4)	38.4 (14.4)	36.8 (11.8)
	Cond. correct recall	55.3 (14.8)	48.3 (14.8)	44.6 (16.5)	39.4 (12.6)
	Recall errors				
	Semantic	27.4 (8.4)	24.4 (11.1)	25.9 (10.9)	25.5 (10.4)
	Phonological	1.9 (2.2)	1.8 (1.9)	3.0 (2.9)	3.2 (3.1)
	Intrusion	6.2 (8.6)	8.3 (12.8)	9.1 (15.5)	8.3 (9.7)
	Other	20.6 (11.6)	22 (14.6)	23.6 (14.5)	26.2 (10.4)

Note. Standard deviations are in brackets.

included both repeated (correct or incorrect) responses and recalled words that were neither related in meaning nor sound to those in the memory list nor originated from another list. Table 2 shows the percentage of recall errors (among the four words to be remembered) by error type as a function of experimental conditions.

A first Bayesian repeated-measures ANOVA[3] compared the rate of each type of incorrect responses (semantic, phonological, intrusions and other errors). The analysis provided decisive evidence for different error rates ($BF_{10} = 1.57 \times 10^{43}$). As expected, post-hoc comparisons indicated that incorrect recall of semantic related words (9.3%, $SD = 7.5$) was more frequent than phonological errors (1.5%, $SD = 2{,}1$; $BF_{10} = 2.11 \times 10^{19}$), intrusions (2.7%, $SD = 4.6$; $BF_{10} = 7.07 \times 10^{12}$) and other errors (2.1%, $SD = 3.5$; $BF_{10} = 5.39 \times 10^{18}$). This experiment thus replicated the findings of Atkins and Reuter-Lorenz (2008), showing that semantic intrusions were prevalent after a brief retention interval of a few seconds, with only a few semantically related words to remember.

As the rate of each type of non-semantic error taken separately was low, we aggregated them for the following analyses. A second Bayesian repeated-measures ANOVA was conducted on the percentage of incorrect responses with attentional demand of the concurrent task, the presence or not of articulatory suppression, and type of errors (semantic vs. non semantic errors) as within-subject factors. Error type was included in the analysis to test the hypothesis that manipulations of WM maintenance mechanisms, particularly articulatory rehearsal, specifically affect short-term semantic errors. To test this hypothesis, it is important to show that our manipulations have less or no effect (or a different effect) on other types of errors. The best model included the main effects of attentional demand, articulatory suppression and error type, and the interaction between articulatory suppression and error type ($BF_{10} = 2.47 \times 10^{19}$). Note that the main effects of attentional demand and articulatory suppression on recall errors were corollaries of their effects on correct recall; performing a concurrent task that was either highly attentionally demanding or performing it under articulatory suppression increased recall error rates. As indicated above, semantic errors were more frequent than non-semantic errors.

To decompose the interaction between articulatory suppression and error type, we conducted Bayesian paired samples *t*-tests separately for each type of errors with articulatory suppression as within-subject factor. As expected, participants made decisively more semantic recall errors when they performed the concurrent task aloud rather than silently ($BF_{10} = 8.39 \times 10^{6}$). By contrast, there was weak evidence for an effect of articulatory suppression on the rate of non-semantic errors ($BF_{10} = 1.91$). These results are consistent with those of other experiments using short-term recognition tests (Abadie & Camos, 2019; Atkins et al., 2011; Macé & Caza, 2011). They suggest that articulatory rehearsal prevents the occurrence of semantic recall errors in short-term memory tests without affecting non-semantic errors. The increase in semantic errors under articulatory suppression could also result from the prevalence of refreshing during maintenance. However, the results showed that there was no interaction between the concurrent attentional demand and error type ($BF_{incl} = 0.18$). As reported above, all error types, semantic and non-semantic, decreased when participants had more opportunities to refresh the words to be memorized.

2.2.3. Recall performance in the delayed test

2.2.3.1. Correct recall. To test our hypotheses on the effects of refreshing and rehearsal on correct recall in the delayed test (Table 1), the same Bayesian repeated-measures ANOVA as for the immediate test was performed on correct recall in the delayed test. The null model was the best, indicating that neither attentional demand of the concurrent task nor articulatory suppression impacted recall performance in the delayed test (Table 2).

To account for the fact that the words to be recalled in the delayed test have already been tested in the immediate test, we computed a delayed recall score conditional on immediate recall (CR_{cond}). This score corresponds to the number of words correctly recalled in the delayed test that were also correctly recalled in the immediate test (DI) divided by the number of words correctly recalled in the immediate test (I).[4]

$$\mathrm{CRcond} = \frac{\sum_{i=1}^{n}\left(\sum_{j=1}^{k}\frac{DI\ i.j}{I\ i.j}\right)}{n} \quad (1)$$

where i denotes each participant and j each trial.

We performed the Bayesian repeated-measures ANOVA on this score. Interestingly, the additive model including the main effects of attentional demand and articulatory suppression was the best ($BF_{10} = 19.2$), but it was preferred to the model including only the main effect of articulatory suppression by a BF_{10} of 0.78, which provided no substantial evidence for the effect of attentional demand. Participants correctly recalled more words when they performed the concurrent task aloud (40%, $SD = 10.1$) rather than silently (33.9%, $SD = 11.2$). This result

[3] The data on recall errors and subjective judgments (see Exp. 3 and 4) were also analyzed using Generalized Linear Mixed Models, which account for the fact that some observations are not independent, such as the different types of recall errors or subjective judgments. The results available on the OSF are similar to those obtained with the Bayesian ANOVAs.

[4] Note that all the words recalled correctly in the delayed test were also recalled correctly in the immediate test. This was the case for all participants in all experiments.

shows that, compared to the words participants had recalled in the immediate test, they forgot more in the delayed test when they had the opportunity to rehearse them. This finding is consistent with studies showing that rehearsal has no positive effect on long-term information retention (e.g., Camos & Portrat, 2015).

2.2.3.2. Recall errors. The percentage of recall errors in the delayed test (out of the forty words to be remembered) by type of errors as a function of experimental condition is shown in Table 2. A first analysis provided decisive evidence of differences in error rates according to type ($BF_{10} = 8.87 \times 10^{22}$). Post-hoc comparisons indicated that, as expected, incorrect recall of semantic related words (5.2%, $SD = 4.3$) was decisively more frequent than phonological errors (0.4%, $SD = 1.0$; $BF_{10} = 8.87 \times 10^{22}$), intrusions (1.5%, $SD = 4.0$; $BF_{10} = 1.04 \times 10^{11}$) and other errors (1.0%, $SD = 2.4$; $BF_{10} = 3.71 \times 10^{16}$).

Next, as in the immediate test, we aggregated the non-semantic error rates and performed the analysis with the three factors. The best model included only the main effect of error type ($BF_{10} = 1.01 \times 10^4$); semantic errors being more frequent than non-semantic errors.

Among the incorrect responses in the delayed test, some had already been generated in the immediate test while others appeared only in the delayed test ("pure" errors of the delayed test). We tested the effect of attentional demand and articulatory suppression on semantic and non-semantic recall errors occurring only in the delayed test. We computed a conditionalized delayed recall error score for semantic and non-semantic errors (E_{cond}). This score corresponds to the number of semantic (or non-semantic) errors produced in the delayed test that were not produced in the immediate test (DD) divided by the total number of semantic (or non-semantic) errors produced in the delayed test (D).

$$\text{Econd} = \frac{\sum_{i=1}^{n}\left(\sum_{j=1}^{k}\frac{DD_{ij}}{D_{ij}}\right)}{n} \quad (2)$$

where i denotes each participant and j each trial.

The percentage of conditionalized semantic and non-semantic recall errors is shown in Fig. 3. The best model included only the main effect of error type ($BF_{10} = 157$), but it was not substantially preferred to the model also including the main effect of articulatory suppression ($BF_{10} = 86.4$) and the one with the interaction between attentional demand and error type ($BF_{10} = 64.9$). Both error types were more frequent when the concurrent task was performed silently (49.4%, $SD = 27.9$) rather aloud (30.7%, $SD = 20.7$), which is probably a corollary of the effect of articulatory suppression on delayed correct recall. To decompose the interaction between attentional demand and error type, Bayesian paired samples *t*-tests were conducted separately for each type of errors with attentional demand as within-subject factor. There was weak evidence that, among semantic errors, those generated only in the delayed test were more frequent when participants performed the low (65.5%, $SD = 25.4$) than the high attentional demand concurrent task (55.8%, $SD = 27.8$; $BF_{10} = 2.02$). By contrast, there was strong evidence against an effect of attentional demand on the non-semantic errors generated only in the delayed test ($BF_{10} = 0.07$). Although the effect is small, due to the low number of recall errors, this last result suggests that increasing refreshing opportunities specifically promotes the emergence of semantic false memories that appear only in the long term. This result is consistent with those obtained previously in recognition tasks (e.g., Abadie & Camos, 2019).

2.2.4. Summary of Experiment 1 results

As predicted, the results of Experiment 1 showed that hindering the use of rehearsal specifically increased the occurrence of semantic errors in the immediate test. Consequently, using this mechanism could prevent the occurrence of these errors in the short term. In contrast, conditionalized delayed semantic errors decreased when the concurrent task was more attentionally demanding than when it was less demanding. This suggests that the more opportunities participants had to refresh the to-be-remembered words, the more semantic errors appeared in the delayed test. Furthermore, correct recall was higher when both mechanisms were available without the two interacting with each other. Unexpectedly, however, correct recall in the delayed test was not impacted by the manipulation of attentional demand in the concurrent task.

3. Experiment 2

The purpose of Experiment 2 was to replicate the findings of Experiment 1 in the immediate recall test, namely the specific role of rehearsal in preventing the incidence of short-term false memories. In addition, the interesting results from Experiment 1 regarding the role of

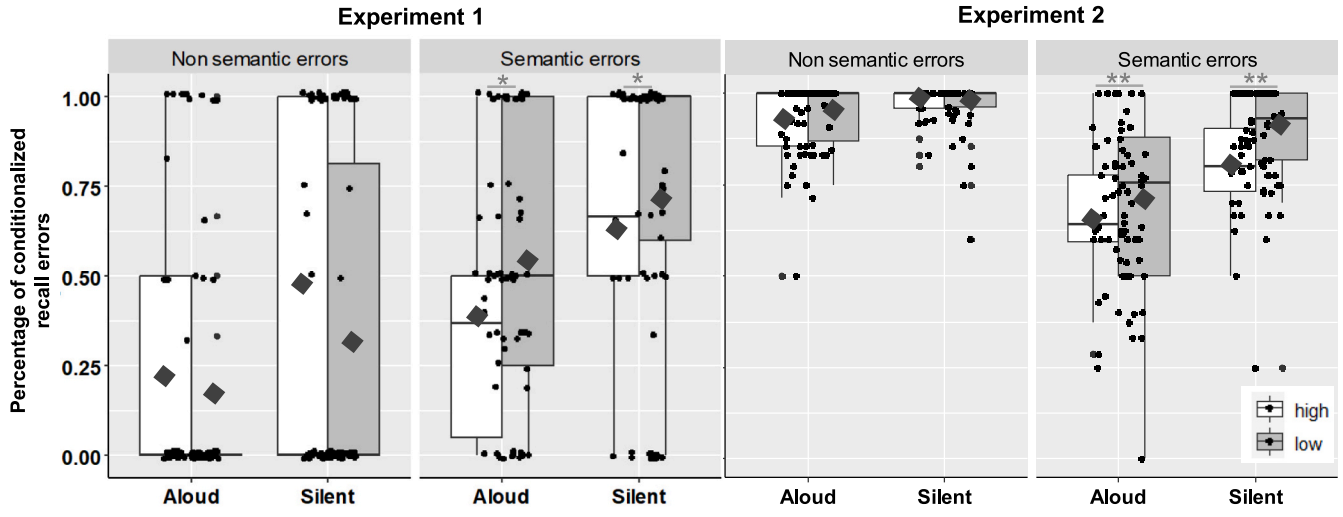

Fig. 3. Percentage of conditionalized semantic and non-semantic errors as a function of attentional demand of the concurrent task (high vs. low) and the presence or not of articulatory suppression (concurrent task performed aloud vs. silently) in Experiments 1 and 2.
Note. In each box plot, the black bar represents the median, the edges of the rectangle are the quartiles, and the ends of the whiskers are calculated using 1.5 times the interquartile range (i.e., the distance between the first and third quartiles). The gray dot represents the mean. Black dots are individual data. ** represents substantial differences between conditions (i.e. Bayes factor > 3). * represents weak differences between conditions (i.e., Bayes factor between 1 and 3) Only the differences predicted in the hypotheses are shown in the figure.

refreshing on long-term false memories merit further examination because the effect was relatively small. In Experiment 1, participants recalled only 16 words on average out of the 40 in the delayed test. This contrasted with the very high recall performance in immediate test (84.5% on average). It should be noted, however, that participants were asked to recall all four words on each trial in the immediate test, whereas they were free to recall as many words as they wished in the delayed test. Therefore, the effects observed in the delayed test, which were rather weak, may be due to participants' low engagement in this test. One of the most common methods used in the LTM literature to encourage participants to take longer to search for information in memory is to force them to recall all the words presented. Numerous studies have shown that the use of a forced recall procedure favors the use of reconstruction processes and the occurrence of recall errors (e.g., Payne et al., 1996; Roediger & Payne, 1985; see Roediger et al., 1993, for a review). This procedure therefore seemed relevant for assessing the effects of manipulating WM maintenance mechanisms on recall errors in the delayed test, distinguishing between their semantic and non-semantic types. In Experiment 2 (and subsequent experiments), participants were then instructed to recall 40 words in the delayed test.

3.1. Method

3.1.1. Participants

Forty Aix-Marseille University students (M_{age} = 20.5 years, SD_{age} = 3.18, range = 18–33 years old; 37 females, 3 males) took part in Experiment 2. They received course credits in return for their participation. All participants were healthy, presented no learning disabilities and were native French speakers. None of them participated to Experiment 1.

3.1.2. Materials and procedure

The same procedure as in Experiment 1 was used (Fig. 2), except that the participants were instructed to recall 40 words in each condition during delayed recall test.

3.2. Results and discussion

As in Experiment 1, the experimenter checked that participants complied with the instructions and paid sufficient attention to the concurrent tasks. Data from four participants were excluded, because their recall performance differed from the average performance by more than 3SDs.

3.2.1. Performance on the mathematical verification task

Performance on the concurrent tasks was good. There was no difference between the conditions in which participants had to verify the mathematical expression aloud (85.8%, *SD* = 16.5) or in silence (85.0%, *SD* = 15.2; BF_{10} = 0.19).

3.2.2. Recall performance in the immediate test

3.2.2.1. Correct recall. The same Bayesian repeated-measures ANOVA as in Experiment 1 was performed on correct recall. As in Experiment 1, the additive model including the main effect of attentional demand and of articulatory suppression was the best ($BF_{10} = 3.15 \times 10^{17}$). As expected, there was decisive evidence that immediate recall performance was better when participants performed the low demanding (86.6%, *SD* = 9.1) than the high demanding (81.5%, *SD* = 10.1) concurrent task and when they performed the concurrent task silently (89.6%, *SD* = 8.2) rather than aloud (78.5%, *SD* = 8.1). Thus, once again, our manipulations designed to affect either refreshing opportunities or rehearsal did reduce immediate recall performance, without interacting with each other.

3.2.2.2. Recall errors. As in Experiment 1, incorrect responses were classified by the same two independent and trained raters into semantic, phonological, intrusions and other errors. Interrater agreement was 76.3% before discussion among raters and full interrater agreement was achieved after discussion. A first Bayesian analysis conducted to compare the rates of the different types of errors provided decisive evidence that they did differ from each other ($BF_{10} = 1.41 \times 10^{60}$). As expected, semantic recall errors (10.4%, *SD* = 7.4) were more frequent than phonological (1.1%, *SD* = 1.6; $BF_{10} = 1.61 \times 10^{27}$), intrusions (2.3%, *SD* = 3.8; $BF_{10} = 6.52 \times 10^{18}$) and other errors (2.2%, *SD* = 3.8; $BF_{10} = 5.64 \times 10^{20}$).

Then, as in Experiment 1, the non-semantic error rates were aggregated, and a Bayesian repeated-measures analysis was conducted with attentional demand, articulatory suppression, and error type as within subject factors. The best model included the main effects of the three factors as well as the interaction between articulatory suppression and error type ($BF_{10} = 3.75 \times 10^{22}$). The decomposition of the interaction between articulatory suppression and error type showed that, as expected, participants made decisively more semantic recall errors when they performed the concurrent task under articulatory suppression than when the task was performed silently ($BF_{10} = 2.05 \times 10^{7}$). By contrast, there was weak evidence for an effect of articulatory suppression on the rate of non-semantic errors (BF_{10} = 1.14). These results, which replicated those of Experiment 1, demonstrated the specific impact of the rehearsal manipulation on semantic errors in immediate recall tests.

3.2.3. Recall performance in the delayed test

All participants complied with the instruction to recall 40 words in each condition.

3.2.3.1. Correct recall. As expected, compared to Experiment 1, forcing participants to recall more memory words led them to better recall performance (31.9%, *SD* = 12.5; 40.6%, *SD* = 8.8, in Exp. 1 and 2 respectively). The analysis revealed that the model including the main effect of attentional demand only was the best (BF_{10} = 25.2). Unexpectedly, recall performance was better in the high (43.7%, *SD* = 12.7) than in the low attention demanding condition (37.6%, *SD* = 13.1).

Next, the same analysis was conducted on the delayed recall score conditionalized on immediate recall (CR_{cond}). The additive model including the main effects of attentional demand and articulatory suppression was the best ($BF_{10} = 7.3 \times 10^{3}$). As with the delayed recall score, more words were correctly recalled in the delayed test in the high (51.8%, *SD* = 14.8) than in the low attention demanding condition (42%, *SD* = 14.5). In addition, as in Experiment 1, conditionalized recall performance was better when participants performed the concurrent task aloud (50%, *SD* = 15.6) rather than silently (43.8%, *SD* = 13.7). These results suggested that neither rehearsal nor refreshing is beneficial to correct recall in the delayed test.

3.2.3.2. Recall errors. As expected, the percentage of the different types of errors increased compared to Experiment 1 as a result of the use of the forced recall procedure. The best model included only the main effect of error type ($BF_{10} = 2.08 \times 10^{76}$). Post-hoc comparisons indicated that semantic recall errors (25.8%, *SD* = 10.2) were more frequent than phonological errors (2.5%, *SD* = 2.7; $BF_{10} = 2.87 \times 10^{53}$) and intrusions (8%, *SD* = 11.9; $BF_{10} = 1.29 \times 10^{18}$), but not more frequent than other errors (23.1%, *SD* = 13.7; BF_{10} = 0.31). Other errors were also more frequent than phonological errors ($BF_{10} = 7.7 \times 10^{34}$) and intrusions ($BF_{10} = 1.06 \times 10^{15}$) and the latter were more frequent than phonological errors ($BF_{10} = 2 \times 10^{4}$). The increase in other errors with this forced recall procedure is simply explained by the fact that participants produced random words when they ran out of words to recall.

Next, as in Experiment 1, we tested the effect of attentional demand and articulatory suppression on semantic and non-semantic (i.e., phonological, intrusion and other error rates being aggregated) recall

errors occurring only in the delayed test (E_{cond}, Fig. 3). Note that the fact that the rate of non-semantic errors is particularly high here simply means that the majority of these errors were produced only in the delayed test as a result of the forced recall procedure, as explained above. The best model included the main effect of the three factors as well as the interaction between articulatory suppression and error type ($BF_{10} = 1.14 \times 10^{32}$), but it was not substantially preferred to the model also including the interaction between attentional demand and error type ($BF_{10} = 1.12 \times 10^{32}$). Follow-up Bayesian repeated-measures ANOVA were conducted separately for each type of errors with attentional demand and articulatory suppression as within subject factors to decompose the interactions. Concerning semantic errors, the additive model including main effects of attentional demand and articulatory suppression was the best ($BF_{10} = 5.23 \times 10^{7}$). As predicted, semantic errors generated only in the delayed test were more frequent in the low (79.3%, *SD* = 16.1) than in the high attentional demand condition (73.9%, *SD* = 11.9). Moreover, as in Experiment 1, semantic errors generated only in the delayed test were more frequent when the concurrent task was performed silently (85.5%, *SD* = 11.8) rather than aloud (67.7%, *SD* = 14.1). Finally, regarding non-semantic errors, the best model included only the main effect of articulatory suppression ($BF_{10} = 5.31$). Again, these errors more frequent when the concurrent task was performed silently (97%, *SD* = 4.3) rather than aloud (93.5%, *SD* = 7.1). As in Experiment 1, and even more so here since participants were required to recall the 40 words, this latter result was simply the corollary of the fact that delayed correct recall was lower in the conditions without concurrent articulation. The effect of the attentional demand manipulation on semantic errors in the delayed test is of particular interest. Although still relatively small, as in Experiment 1, it points in the same direction. This effect suggests that refreshing might specifically promote the occurrence of semantic errors in the long term.

3.2.4. Summary of Experiment 2 results

The findings of Experiment 2 replicated those of Experiment 1. First, correct recall was increased, and semantic errors decreased in the immediate test when participants could use rehearsal. Although reducing refreshing opportunities reduced immediate correct recall, it had no specific effect on immediate semantic errors. Delayed recall errors were more frequent than in Experiment 1 due to forced recall instructions. However, as in Experiment 1, conditionalized semantic errors were more frequent with enhanced opportunities for refreshing. Nevertheless, the effect was still quite small. Finally, conditionalized delayed correct recall was higher when opportunities for refreshing or rehearsal were reduced. This latter result questions the role of these two mechanisms in long-term correct recall.

4. Experiment 3

Experiment 3 was then designed to specifically investigate the role of refreshing on the occurrence of false memories. The use of rehearsal was systematically prevented by having participants continuously utter the syllables "ba-bi-boo" aloud during the retention interval. In addition, rather than asking participants to verify or read a single mathematical expression during the retention interval, we implemented a continuous operation task (e.g., Barrouillet et al., 2004) in which participants were either asked to give the results of several simple additions and subtractions or to simply press a key each time a result appeared on screen (simple reaction time, SRT, task). These tasks allow for a stronger manipulation of the concurrent attentional demand, i.e., the opportunities for refreshing.

In line with what has been shown in the LTM literature, the forced delayed recall procedure introduced in Experiment 2 increased both correct recall of the studied items and recall errors compared to Experiment 1 in which delayed recall was free. However, one criticism of this procedure is that it encourages the emergence of different types of responses (e.g., Roediger et al., 1993). For example, some of the recalled words may have been recalled with a high level of confidence that they were presented during the study phase, while other words may have been recalled with a medium level of confidence and a fuzzier memory of their previous presentation, and still others may have been randomly given. Therefore, it seems important to be able to distinguish between these different types of responses. In Experiment 3, we asked participants to assign a confidence level to each of their responses.

We expected correct responses to be more often associated with high confidence levels in both recall tests. While non-semantic errors should be predominantly associated with low confidence levels, semantic errors should be associated with higher confidence levels than non-semantic errors, but lower than correct responses. Some studies using the classic DRM paradigm have shown that for over 85% of false memories of critical lures, participants reported some confidence that they were in the studied list (e.g., Payne et al., 1996). This phenomenon has been termed phantom recollection (e.g., Brainerd et al., 2001; Lampinen et al., 1998, 2008). When there is a high number of gist-based false memories, a subset of those memories may be accompanied by vivid false recollective experiences. Several studies have also reported the occurrence of short-term false memories accompanied by phantom recollection (46%, Abadie & Rousselle, 2023) or "remember" responses (around 30%, Flegal et al., 2010; Flegal & Reuter-Lorenz, 2014). We also examined the effect of manipulating refreshing opportunities on the subjective experience reported by participants. If, as in Abadie and Camos' (2019) study, the use of refreshing strengthens the retrieval of gist memories in the delayed test, then semantic errors should be associated with higher confidence in the delayed test when the concurrent task is low rather than high demanding. These false memories should also be more often accompanied by moderate confidence when there were more opportunities for refreshing because, on our confidence scale (see method below), moderate confidence also reflects retrieval of the general meaning (or gist) of one or more studied words.

4.1. Method

4.1.1. Participants

Forty-one Aix-Marseille University students (M_{age} = 21.9 years, SD_{age} = 4.35, range = 18–30 years old; 31 females, 10 males) took part voluntarily in Experiment 3. All participants were healthy, presented no learning disabilities and were native French speakers. None of them participated in previous experiments.

4.1.2. Materials

4.1.2.1. Word lists. To ensure that the effects obtained in the first two experiments were not contingent on the word lists used, 20 semantically related lists were selected anew from the verbal association norms for concrete French nouns (Bonin et al., 2013) using the same criteria as in previous experiments. These 20 lists were separated into two groups of 10 four-word lists that were equated in mean BAS (*M* = 0.22 for each group of lists). Each group was assigned in a counterbalanced manner to the two experimental conditions. The order of presentation of the lists was randomized.

4.1.2.2. Concurrent task. We created 38 simple mathematical expressions that were all different. Twenty of them were used in the experimental conditions, 10 per condition, the others were used during the training phase. Each expression consisted of a starting root digit followed by two consecutive operations that could be either additions or subtractions (e.g., "7 -2 +1". In this example "7" is the root digit, "-2" the first operation, "+1" the second operation). Only four types of simple operations were used: "-2", "-1", "+1" and "+2", and the results from these operations were integers ranging from 1 to 9. The root digit, operations and results were each presented sequentially in the center of the screen. To help participants distinguish each part of the mathematical

expression, the root digit was presented in green, the operations in black, and the results in red.

4.1.2.3. Subjective experience. During the recall phases, participants were asked to give a confidence judgment for each recalled word on a 3-point scale ranging from 1, "I'm giving a random answer", to 3, "I'm absolutely sure that the word I just recalled was one of the studied words". Response 2 was to be given when participants were moderately confident and didn't know whether the recalled word was a studied word or a word semantically related to a studied word. There was no time limit for these qualitative judgments.

4.1.3. Procedure

Experiment 3 was performed online due to the COVID 19 epidemic and was then programmed using the online experimental platform Labvanced (Finger et al., 2017). It was launched remotely on the participant's personal computer. Simultaneously, participants were assisted by the experimenter using a video conferencing tool. Their progress was monitored live, and they were in constant communication with the experimenter throughout the study, making the testing conditions very similar to those in Experiments 1 and 2, in which the experimenter was in the room with the participants.

The procedure was similar to that of the previous experiments (Fig. 2). At the beginning of each trial, four-word associates of a given list appeared simultaneously on the screen for 1750 ms. The participants were asked to read the words out loud. Then, during the retention interval, participants were instructed to complete the concurrent task. They first heard a 250 ms beep sound indicating that they should begin to repeat the syllables "ba-bi-boo" continuously. This concurrent articulation, which was intended to impede the use of articulatory rehearsal in both experimental conditions, was continuous and lasted for the full duration of the concurrent task, i.e., 4000 ms. The mathematical task varied depending on the experimental condition. In the high attentional demand condition, the root digit appeared on the screen for 1000 ms, followed by the operations, which remained each on the screen for 1500 ms during which participants had to press the number corresponding to the result of the operation on their keyboard. In contrast, in the low attentional demand condition, each operation was presented for 750 ms on the screen, followed by its result, which also remained on the screen for 750 ms, during which participants were asked to press "0" on their keyboard. Then, participants were prompted to recall the four words (immediate recall). A confidence judgment was to be associated with each recalled word. Recall responses and confidence judgments were given orally to the experimenter. There were ten trials per condition.

At the end of each condition, participants were asked to count backwards by twos from a randomly chosen number between 100 and 1000 for 1 min. They were then instructed to orally recall the 40 words they had just seen in the condition preceding the countdown (delayed recall) and to associate a confidence judgment with each recalled word. There were no time limits on the recall phases. Each participant completed both conditions, the order of which being counterbalanced across participants.

All participants underwent a training phase prior to the experiment. During this phase, the experimenter not only ensured that participants understood the instructions, as in previous experiments, but also that they mastered the use of the confidence judgments, so that they did not interfere with word recall. Finally, demographic information was collected at the end of the experiment. Information regarding participants' gender was collected using a drop-down menu with three options (i.e., female, male and other), and date of birth using an input box for participants to enter.

4.2. Results and discussion

One participant did not follow instructions correctly and five other participants failed to respond (i.e., omissions) to the concurrent task in >10% of trials and were therefore not included in the analyses.

4.2.1. Performance on concurrent tasks

Performance on the concurrent tasks was very good and participants were better on the low demanding task (96.8%, $SD = 6.26$) than on the high demanding task (86.9%, $SD = 11.1$; $BF_{10} = 329$). As expected, they also responded faster to the low- (394 ms, $SD = 52$) than to the high-demanding task (1042 ms, $SD = 74.6$; $BF_{10} = 2.77 \times 10^{27}$).

4.2.2. Recall performance in the immediate test

4.2.2.1. Correct recall. The percentage of correct and incorrect responses associated with each level of confidence is presented in Table 3. To test our hypotheses regarding the effect of refreshing on correct recall as a function of confidence level associated to the response in the immediate test, we performed a Bayesian repeated-measures ANOVA with attentional demand and confidence level as within-subject factors on correct recall. The full model including the main effects of both factors as well as their interaction was the best ($BF_{10} = 1.15 \times 10^{60}$). As in previous experiments, recall performance was better when participants performed the low than the high demanding task, which suggests that correct recall was boosted when there were more opportunities for refreshing in the immediate test. As predicted, post-hoc comparisons showed that correct responses were more often associated with high confidence than with medium ($BF_{10} = 8.06 \times 10^{4}$) or low ($BF_{10} = 1.80 \times 10^{62}$) confidence. Responses associated with medium confidence were also more frequent than random responses ($BF_{10} = 8.09 \times 10^{3}$). To decompose the interaction between attentional demand and confidence level, paired Bayesian *t*-tests were conducted with attentional demand as a within-subject factor for each confidence level separately. The analyses provided substantial evidence of the effect of attentional demand only for responses associated with a high level of confidence ($BF_{10} = 3.32$).

4.2.2.2. Recall errors. As in previous experiments, incorrect responses were classified by independent and trained raters into semantic, phonological, intrusions and other errors. Interrater agreement was 96% and full interrater agreement was achieved after discussion. A first Bayesian repeated-measures analysis was conducted with attentional demand and confidence level as within-subject factors on semantic errors. Although the model including only the main effect of confidence level was the best ($BF_{10} = 3.38 \times 10^{10}$), it was preferred only by a BF_{10} of 1.98 to the additive model, indicating that both confidence level and attentional demand could have an effect. Semantic errors appeared to be slightly more frequent in the high- than the low-demanding condition, but the analysis of effects provided insensitive evidence for the main effect of attentional demand ($BF_{incl.} = 0.56$). Post-hoc comparisons showed that semantic errors were more often associated with low confidence than with high ($BF_{10} = 2.95 \times 10^{7}$) or medium ($BF_{10} = 628$) confidence. Semantic errors associated with medium confidence were also more frequent than those associated with high confidence ($BF_{10} = 6.56$).

For the (aggregated) non-semantic errors, the best model included only the main effect of confidence level ($BF_{10} = 1.22 \times 10^{4}$). Post-hoc comparisons indicated that non-semantic errors were more often associated with low confidence than with medium ($BF_{10} = 50.2$) or high (0.85%, $SD = 1.55$; $BF_{10} = 616$) confidence.

4.2.3. Recall performance in the delayed test

4.2.3.1. Correct recall. As with immediate recall, a first analysis was performed on correct recall with attentional demand and confidence level as within-subject factors. The best model included the main effect of confidence only ($BF_{10} = 4.81 \times 10^{83}$). Post-hoc comparisons showed

Table 3
Mean percentage of responses associated with a confidence judgment on the 3-point scale for correct recall and recall errors (semantic vs. non semantic errors) as a function of concurrent task attentional demand (high vs. low) for immediate and delayed tests in Experiment 3.

	High demanding concurrent task			Low demanding concurrent task		
Recall accuracy	Judgment 1-guess	Judgment 2-medium	Judgment 3-sure	Judgment 1-guess	Judgment 2-medium	Judgment 3-sure
			Immediate test			
Correct recall	2.0 (3)	5.4 (4.9)	79.8 (8.8)	1.1 (1.9)	5.6 (5)	83.1 (8)
Recall errors						
Semantic	5 (4.5)	1.7 (2.7)	0.8 (2)	3.5 (2.9)	1.8 (2.6)	0.4 (1.1)
Phonological	0.3 (0.8)	0.4 (1)	0.6 (1.2)	0.1 (0.6)	0.4 (0.9)	0.2 (0.7)
Intrusion	0.9 (1.5)	0.5 (1)	0.6 (1.5)	0.9 (1.6)	0.8 (1.5)	0.4 (1.1)
Other	1.9 (4)	0.3 (1)	0.07 (0.4)	1.6 (3)	0.07 (0.4)	0.0 (0)
			Delayed test			
Correct recall	2.3 (2.9)	7.1 (5.2)	44.6 (12.1)	1.4 (1.9)	6.9 (7.6)	43.2 (12.4)
Cond. correct recall		60.1 (15)			55.8 (14.3)	
Recall errors						
Semantic	8.8 (6.1)	3.9 (3.3)	1.4 (1.5)	6.8 (5.5)	3.9 (4.1)	1.3 (2)
Phonological	0.1 (0.6)	0.3 (0.8)	0.6 (1.2)	0.0 (0)	0.3 (0.8)	0.1 (0.6)
Intrusion	1.2 (2.9)	0.6 (1.5)	0.8 (2.3)	3.0 (6.5)	1.0 (2.2)	1.5 (3.6)
Other	22.6 (13.9)	2.1 (4.5)	0.1 (0.6)	24.2 (13.7)	2.3 (3.3)	0.5 (1.8)

Note. Standard deviations are in brackets. Judgment 1 is indicative of a mere guess, judgment 2 represents a response with medium confidence, and judgment 3 a response with high confidence.

that correct responses with high confidence were more frequent than correct responses with medium ($BF_{10} = 2.55 \times 10^{28}$) and low ($BF_{10} = 1.82 \times 10^{35}$) confidence. Correct responses with medium confidence were also more frequent than those with low confidence ($BF_{10} = 1.14 \times 10^{6}$). As in previous experiments, the analysis provided substantial evidence against an effect of concurrent attentional demand on correct delayed recall ($BF_{10} = 0.30$).

As in the two first experiments, a second analysis was conducted with attentional demand as a within-subject factor on the delayed recall score conditionalized on immediate recall (CR_{cond}). Again, the analysis provided weak evidence against the effect of concurrent attentional demand ($BF_{10} = 0.49$). Participants correctly recalled over 50% of the words to be remembered, which is slightly higher than in the two first experiments, but there was still no evidence of a beneficial effect of refreshing on long-term retention.

4.2.3.2. Recall errors. To test our hypotheses regarding the effect of refreshing on semantic errors as a function of confidence level in the delayed test, a Bayesian repeated-measures analysis with attentional demand and confidence level as within-subject factors was performed on delayed semantic errors. The analysis indicated that the best model included only the main effect of confidence level ($BF_{10} = 1.93\times 10^{14}$). Post-hoc comparisons indicated that semantic errors were more often associated with low confidence judgments than with medium ($BF_{10} = 413$) or high confidence judgments ($BF_{10} = 3.09 \times 10^{9}$). Semantic errors associated with medium confidence were also more frequent than semantic errors associated with high confidence ($BF_{10} = 8494$).

For non-semantic errors, the model including only the main effect of confidence was also the best ($BF_{10} = 6.95 \times 10^{42}$). Post-hoc comparisons showed that non-semantic errors associated with low confidence were more frequent than non-semantic errors with medium confidence ($BF_{10} = 4.91 \times 10^{16}$) or high confidence ($BF_{10} = 1.04 \times 10^{18}$). Non-semantic errors associated with medium confidence were also more frequent than non-semantic errors with high confidence ($BF_{10} = 8.17$).

Next, as in previous experiments, we tested the effect of attentional demand on semantic and non-semantic recall errors occurring only in the delayed test (E_{cond}). We could not consider the confidence level in these analyses, as it could vary between the immediate and delayed tests for the same semantic error. For semantic errors (Fig. 4), a Bayesian paired samples t-test provided weak evidence that they were more frequent in the low than in the high demanding condition ($BF_{10} = 1.22$). In contrast, the analysis provided substantial evidence against the effect of attentional demand on non-semantic errors generated only in the delayed test ($BF_{10} = 0.11$).

4.2.4. Summary of Experiment 3 results

As expected, in the immediate test, correct responses were more often associated with high confidence. They were more numerous when the attentional demand of the concurrent task was low, i.e., when participants had more opportunities to refresh the memory words. Correct recall performance in the delayed test was also more often associated with a high level of confidence. However, as in previous experiments, attentional demand had no effect on delayed correct recall. As expected, there was no effect of attentional demand on semantic errors in the immediate test. However, these errors were indistinguishable from non-semantic errors in terms of the level of confidence associated with them. Semantic errors appearing only in the delayed test were more frequent in the low attentional demand condition (although the effect was small). Finally, semantic errors made in the delayed test were associated with higher confidence levels than non-semantic errors. Forty percent of the semantic errors were associated with moderate and high confidence levels, whereas only 16.6% of the non-semantic errors were associated with these confidence levels.

5. Experiment 4

The first goal of Experiment 4 was to implement an even stronger manipulation of the attentional demand of the concurrent task by varying the pace of digits (either slow or fast) during the retention interval to ensure that its effects on long-term recall were not related to an insufficiency strong manipulation. It has been extensively shown that increasing the pace of presentation of distractors in a WM task has a detrimental effect on recall performance, because it induces a strong capture of the attention needed for memory maintenance (e.g., Barrouillet et al., 2007; see Barrouillet & Camos, 2015, for a review).

Our second goal was to improve the rating scale for participants' subjective experience. In Experiment 3, we used confidence judgments. Semantic errors were associated with higher confidence than non-semantic errors, but only in the delayed test, and there was no interaction between confidence level and attentional demand of the concurrent task in the delayed test. However, it is possible that asking for the confidence level associated with each response did not really capture the retrieval of gist representations. Tulving (1985) has shown that participants are able to consciously distinguish between memories for which they can recollect vivid details of the event ("remember") and those for which they have a fuzzier memory ("know"). The famous remember/know paradigm he developed has since been used in hundreds of studies to dissociate these two types of memories in recall and recognition tasks (see Yonelinas, 2002, for a review). In Experiment 4, we adapted this

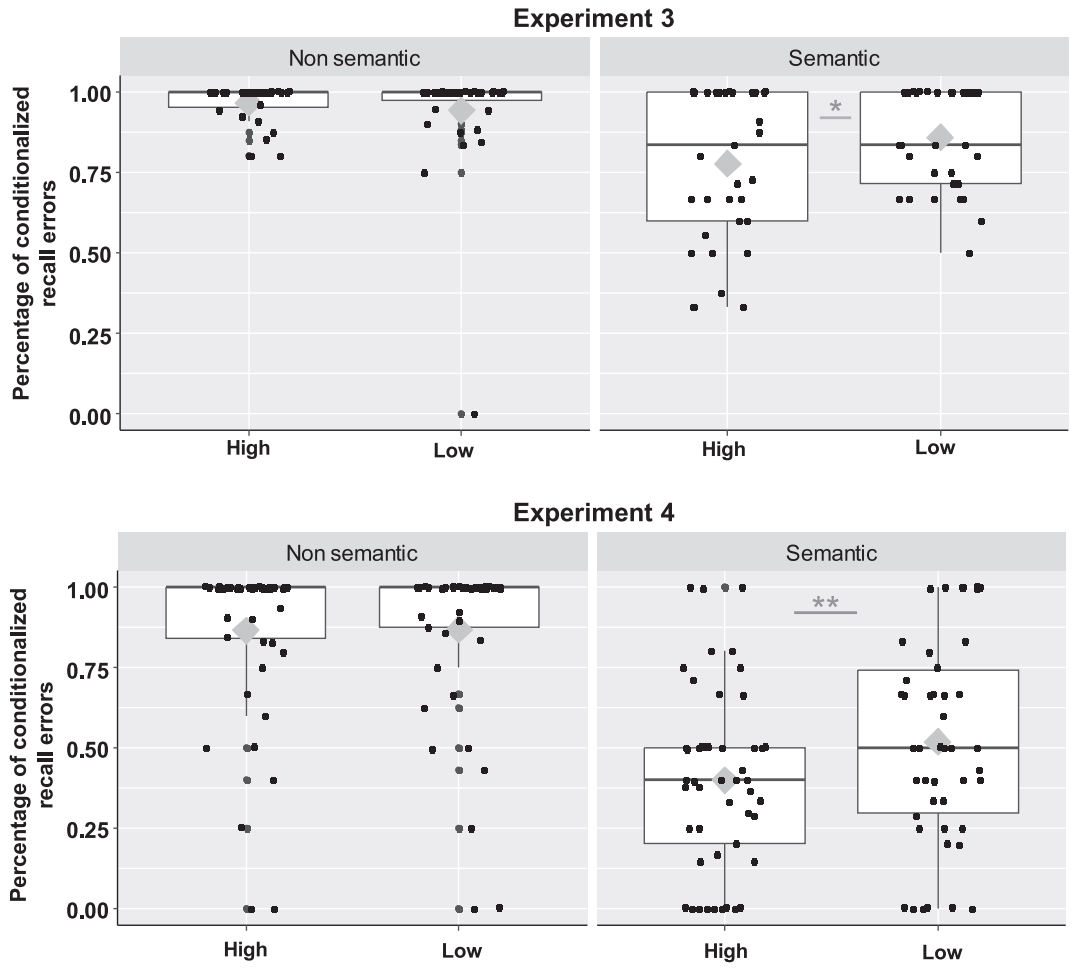

Fig. 4. Percentage of conditionalized semantic and non-semantic errors as a function of attentional demand of the concurrent task (high vs. low) in Experiments 3 and 4.

Note. In each box plot, the black bar represents the median, the edges of the rectangle are the quartiles, and the ends of the whiskers are calculated using 1.5 times the interquartile range (i.e., the distance between the first and third quartiles). The gray dot represents the mean. Black dots are individual data. ** represents substantial differences between conditions (i.e. Bayes factor > 3). * represents weak differences between conditions (i.e., Bayes factor between 1 and 3) Only the differences predicted in the hypotheses are shown in the fig.

paradigm to our task, based on the distinction between verbatim and gist representations proposed by the FTT (Reyna & Brainerd, 1995). Participants were asked to indicate for each recalled word whether they thought it was a word from the study list for which they could retrieve some details of its presentation (a "studied" response, presumably based on verbatim memory retrieval), whether they had a vague memory of its presentation and did not know whether this word or another semantically related word had been presented (a "studied or related" response, presumably based on gist memory retrieval) or whether they had given this word at random (a "guess" response). With this new procedure, we expected that semantic errors would be associated with "studied or related" and "studied" responses more often than other error types in both tests. In addition, the effect of the attentional demand manipulation would have more chances to appear on these semantic errors in the delayed test.

5.1. Method

5.1.1. Participants

Forty-nine young adults ($M_{age} = 19.6$ years, $SD_{age} = 0.85$, range = 18–22 years old, 38 females, 10 males and 1 other gendered person) took part in Experiment 4. All participants were native-French-speaker students at Aix-Marseille University. They were all healthy, had no learning disabilities and had not participated in previous experiments.

5.1.2. Materials

5.1.2.1. Word lists. From the set used in Experiment 3, we selected 12 semantically related lists that resulted in the higher rate of semantic errors. These lists were separated into two groups of 6 four-word lists that were equated in mean BAS ($M = 0.22$ for each group of lists). As in Experiment 3, each group was assigned in a counterbalanced manner to the two experimental conditions. The order of presentation of the lists was randomized.

5.1.2.2. Concurrent task. Twenty-two new mathematical expressions were created in the same manner as in Experiment 3. Twelve were used in the experiment and 10 in the training phase. Each expression started with a root digit followed by 3 consecutive operations in the low attentional demand condition and 5 operations in the high attentional demand condition. Contrary to Experiment 3, the results of the operations were never displayed. To help participants distinguish between each part of the mathematical expression, the root digit was framed by a black square and the operations were each presented in a different color.

5.1.2.3. Subjective experience. For each recalled word, participants were asked to indicate whether they thought it was one of the studied words (i.e., a "studied" response), whether they were familiar with the gist or

the meaning of the recalled word but were unsure whether it was a studied word or a word semantically related to one of the studied words (i.e., a "studied or related" response) or whether they had recalled a word at random (i.e., a "guess" response). This scale, inspired by the study of Brainerd et al. (2010), should make it possible to grasp the level of precision of the memory.

5.1.3. 5.1.3. Procedure

Like Experiment 3, this experiment was programmed using the Labvanced online experimental platform (Finger et al., 2007) and conducted online due to the COVID 19 epidemic. The procedure was similar to that of Experiment 3, except there were only six trials per condition and the pace of the concurrent task was varied rather than the nature of the task (Fig. 2). The root digit was presented for 500 ms. In the high demanding condition, the five operations were displayed sequentially for 900 ms each. First, participants had to calculate the result of the root digit ("2", for instance) and operation 1 (e.g., "-1"), then of the first result (e.g., "1") and operation 2 (e.g., "+2"), then of the second result (e.g., "3") and operation 3 (e.g., +1), etc. The same procedure was used in the low demanding condition, except that only 3 operations were displayed sequentially for 1500 ms each. Articulatory rehearsal was blocked in both conditions with participants being prompted to read the root digit and give the results out loud.

5.2. Results and discussion

Data from three participants were excluded because their recall performance differed from the average performance by more than 3SDs.

5.2.1. Performance on concurrent tasks

Performance on the concurrent tasks was very good. As in Experiment 3, participants performed better on the low demanding task (94.6%, $SD = 9.17$) than on the high demanding task (81.3%, $SD = 14.9$; $BF_{10} = 1.84 \times 10^{5}$).

5.2.2. Recall performance in the immediate test

5.2.2.1. Correct recall. As in Experiment 3, a first analysis was performed on correct recall with attentional demand and judgment type as within-subject factors. The best model included the main effect of judgment type only ($BF_{10} = 3.61 \times 10^{45}$) but was not substantially preferred to the second-best model that also included the main effect of attentional demand ($BF_{10} = 2.5 \times 10^{45}$). As expected, recall performance was better when participants performed the low than the high demanding task (Table 4). Post-hoc comparisons indicated that, as predicted, correct recall was more often accompanied by a "studied" judgment than by a "studied or related" ($BF_{10} = 3.82 \times 10^{34}$) or a "guess" judgment ($BF_{10} = 3.86 \times 10^{52}$). "Studied or related" judgments were also more frequent than "guess" judgments ($BF_{10} = 8.13 \times 10^{15}$).

5.2.2.2. Recall errors. As in the previous experiments, incorrect responses were classified by independent and trained raters into semantic, phonological, intrusions and other errors (Table 4). Interrater agreement was 97.5% and full interrater agreement was achieved after discussion. The Bayesian analysis with attentional demand and subjective experience on semantic errors showed that, although the best model included only the main effect of judgment type ($BF_{10} = 8.68 \times 10^{11}$), it was not substantially preferred to the additive model, which was the second best ($BF_{10} = 4.88 \times 10^{11}$). Semantic errors were more frequent in the high (vs. low) demand condition, but paired *t*-tests indicated that the effect of attentional demand was weak and present only for semantic errors accompanied by "studied or related" judgment ($BF_{10} = 2.21$). Semantic errors were more often accompanied by "studied or related" or "guess" judgments than by "studied" judgments ($BF_{10} = 1.81 \times 10^{11}$; $BF_{10} = 2.41 \times 10^{8}$, respectively). There was substantial evidence that the rates of the first two judgments did not differ ($BF_{10} = 0.16$).

Finally, regarding non-semantic errors, the model including only the main effect of judgment type was the best ($BF_{10} = 2.6 \times 10^{6}$). As expected, non-semantic errors were more often associated with "guess" judgments than "studied or related" ($BF_{10} = 95.9$) or "studied" judgments ($BF_{10} = 4244$). "Studied or related" judgments were also more frequently associated with non-semantic errors than "studied" judgments ($BF_{10} = 3.33$).

5.2.3. Recall performance in the delayed test

5.2.3.1. Correct recall. For correct recall, the analysis showed that, although the best model included only the main effect of judgment type ($BF_{10} = 3.1 \times 10^{92}$), it was not substantially preferred to the second-best model that also included the main effect of attentional demand and the interaction between both factors ($BF_{10} = 1.34 \times 10^{92}$). As shown in Table 4, correct recall was better in the low attentional demand condition. Post-hoc comparisons indicated that correct recall was more often accompanied by "studied" (42.2%, $SD = 14.4$) than "studied or related" (8.08%, $SD = 7.61$; $BF_{10} = 1.12 \times 10^{30}$) and "guess" judgments (1.79%, $SD = 2.74$; $BF_{10} = 2.16 \times 10^{41}$). There was also decisive evidence that "studied or related" judgments were more frequent than "guess" judgments ($BF_{10} = 2.56 \times 10^{8}$). Follow-up paired *t*-tests provided weak evidence that, when correct recall was associated with "studied" judgments, recall performance was better in the low attentional demand

Table 4
Mean percentage of responses associated with a "guess", "studied or related", or "studied" judgment for correct recall and recall errors (semantic vs. non-semantic errors) as a function of concurrent task attentional demand (high vs. low) for immediate and delayed tests in Experiment 4.

	High demanding concurrent task			Low demanding concurrent task		
Recall accuracy	"Guess" judgment	"Studied or related" judgment	"Studied" judgment	"Guess" judgment	"Studied or related" judgment	"Studied" judgment
			Immediate test			
Correct recall	2.4 (3.4)	13.8 (9)	57.8 (14.2)	2.6 (4)	14.6 (9.5)	61.3 (12.6)
Recall errors						
Semantic	9.4 (9.8)	8.8 (7.6)	1.4 (2.9)	8.1 (7.8)	6.4 (5.7)	1.0 (2)
Phonological	0.09 (0.6)	0.5 (1.4)	0.4 (1.2)	0.09 (0.6)	0.5 (1.7)	0.3 (1)
Intrusion	0.4 (1.2)	0.5 (1.4)	0.2 (0.9)	0.9 (2.9)	0.6 (1.7)	0.2 (0.9)
Other	4.3 (7.2)	0.09 (0.6)	0.0 (0)	2.8 (5.6)	0.3 (1.4)	0.2 (0.9)
			Delayed test			
Correct recall	1.9 (2.9)	8.7 (7.8)	40.5 (13.3)	1.4 (2.5)	7.7 (7.7)	46 (14.5)
Cond. correct recall		68.4 (16.7)			68.2 (16.3)	
Recall errors						
Semantic	8.5 (8.3)	8.3 (7)	3.6 (3.5)	6.8 (7.5)	7.2 (6.3)	2.8 (4)
Phonological	0.09 (0.6)	0.09 (0.6)	0.4 (1.2)	0.0 (0)	0.3 (1)	0.5 (1.3)
Intrusion	2.4 (6.5)	2.0 (4.5)	4.1 (7.3)	2.8 (7.1)	1.4 (3.5)	3.6 (6.1)
Other	13.2 (13.7)	1.8 (4.5)	0.09 (0.6)	14.9 (14.5)	1.1 (2.2)	0.09 (0.6)

Note. Standard deviations are in brackets.

(46%, $SD = 14.5$) than in the high attentional demand condition (40.5%, $SD = 13.3$; $BF_{10} = 2.05$), whereas there was substantial evidence against an effect of attentional demand when correct recall was associated with "studied or related" and "guessing" judgments ($BF_{10} = 0.22$; $BF_{10} = 0.31$, respectively).

Next, an analysis was conducted on the delayed recall score conditionalized on immediate recall (CR_{cond}). Contrary to what was observed for correct recall, the analysis provided substantial evidence against an effect of attentional demand ($BF_{10} = 0.16$). Thus, the effect of attentional demand disappeared when controlling for test-retest. This suggests that the effect of refreshing opportunities on delayed correct recall mirrors its effect on the immediate test.

5.2.3.2. Recall errors. For semantic errors, although the best model included only the main effect of judgment type ($BF_{10} = 5.49 \times 10^4$), it was preferred only by a BF_{10} of 2.10 to the second-best model that also included the main effect of attentional demand. As in the immediate test, semantic errors were more often accompanied by "studied or related" and "guess" judgments than "studied" judgments ($BF_{10} = 8.33 \times 10^4$; $BF_{10} = 867$, respectively). There was substantial evidence that the rate of the two first judgments did not differ ($BF_{10} = 0.12$). Although semantic errors seemed to be more frequent in the high attentional demand than in the low demand condition, paired *t*-tests comparisons provided substantial evidence against an effect of attentional demand on semantic errors associated with each judgment type ($BF_{10} = 0.31$; $BF_{10} = 0.29$; $BF_{10} = 0.28$, for "studied", "studied or related" and "guess" judgments, respectively).

For non-semantic errors, the model including the main effect of judgment type only was the best ($BF_{10} = 9.07 \times 10^{16}$). Post-hoc comparisons indicated that non-semantic errors were more often accompanied by "guess" judgments than by "studied or related" ($BF_{10} = 2 \times 10^8$) and "studied" judgments ($BF_{10} = 5.91 \times 10^6$). There was no difference between the rates of the latter two types of judgment ($BF_{10} = 0.13$).

Next, as in previous experiments, we tested the effect of attentional demand on semantic and non-semantic errors occurring only in the delayed test (Fig. 4). The additive model was the best ($BF_{10} = 1.83 \times 10^{15}$), but it was preferred only by a BF_{10} of 1.18 to the full model that also included the interaction between attentional demand and error type. Therefore, we conducted Bayesian paired t-tests for semantic and non-semantic errors separately with attentional demand as a within-subject factor. There was substantial evidence that semantic errors generated in the delayed test only were more frequent in the low attentional demand (51.1%, $SD = 30.1$) than in the high attentional demand condition (38.7%, $SD = 27.4$; $BF_{10} = 3.78$). By contrast, there was substantial evidence against an effect of attentional demand for non-semantic errors ($BF_{10} = 0.18$). As in previous experiments, semantic errors generated in the delayed test only were increased in the low attentional demand condition. Using a more stringent manipulation of attentional demand, we were able to obtain a larger effect here than in previous experiments.

5.2.4. Summary of Experiment 4 results

Immediate recall performance increased when participants had more opportunities to refresh the memory words. As in previous experiments, manipulating the attentional demand of the concurrent task had no effect on delayed correct recall. Although the percentage of correct recall was higher in the low attentional demand condition, the conditionalized score showed that this effect rather reflected the influence of immediate recall on delayed recall. Most of the time when they recalled a word correctly in the immediate or the delayed test, participants reported remembering details of the word presentation during the study phase (78.1% and 81.4%, respectively). As expected, the more stringent manipulation of attentional demand increased its effect on semantic errors made only in the delayed test. The latter were more numerous in the low attentional demand condition. Interestingly, the majority of semantic errors in the delayed test (and half of them in the immediate test) were associated with either a sense of recollecting details of the presentation of the words falsely recalled (17.2% and 6.8% in the delayed and the immediate test, respectively) or with a vague memory of the meaning or gist of the falsely recalled words (41.8% and 43.1% in the delayed and the immediate test, respectively), while non-semantic errors were mainly attributed to chance (68.3% and 69.1% in the delayed and the immediate test, respectively).

6. General discussion

In studies of false memories, there has been considerable interest in recent years in whether reconstructive retrieval processes that give rise to false memories in LTM are also active in WM. This question echoes the broader question of the relationships between LTM and WM. The apparent convergence of false memory phenomena occurring in short- and long-term tests provides prima facie support for models that describe WM as the activated part of LTM, and contradicts the alternative view that WM is a separate system from LTM. Abadie and Camos (2019) recently proposed a model that accounts for short- and long-term false memories, attempting to integrate WM and LTM approaches. Evidence from recognition tasks in adults and children supports this integrative model (Abadie & Camos, 2019; Abadie & Rousselle, 2023; Rousselle et al., 2022). However, questions have been raised about the adequacy of using recognition tasks to assess the involvement of WM processes, as recall tasks seem more appropriate (e.g., Allen et al., 2018; Malmberg, 2008; Uittenhove et al., 2019). Unfortunately, most of the previous studies on short-term false memory have used recognition tasks (e.g., Abadie & Camos, 2019; Atkins & Reuter-Lorenz, 2008, 2011; Coane et al., 2007; Flegal et al., 2010; 2014) and very few have used recall tasks (to our knowledge, only Atkins & Reuter-Lorenz, 2008).

Our series of experiments made three major contributions. First, they provided new evidence that false memories can occur in WM recall tasks. Second, they were the first to examine the role of active WM maintenance in the false memory phenomenon using recall tasks. In the immediate test, semantic errors were on average three times more frequent when maintenance in WM was impeded by a concurrent task than when it was not. In contrast, in the delayed test, when only the errors produced in this test were considered, the semantic error rate decreased when maintenance in WM was impeded. The third contribution of our study was to dissociate the role of the two main WM maintenance mechanisms, articulatory rehearsal and attentional refreshing (e.g., Baddeley, 1986; Camos et al., 2018), on the formation of false memories. Results from all four experiments consistently showed that semantic errors in the immediate test were reduced when participants had the opportunity to rehearse memory words. In contrast, semantic errors occurring only in the delayed test were increased when refreshing opportunities were provided. These findings replicate those obtained in recognition tasks (e.g., Abadie & Camos, 2019), supporting the idea that each WM maintenance mechanism differentially moderates the occurrence of short- and long-term false memories.

In the following, we consider the contribution of our results to the understanding of short-term false memories, the role of articulatory rehearsal on the occurrence of these errors in the short term, and the role of attentional refreshing on their longer-term occurrence. Finally, we consider the question of shared processes between WM and LTM.

6.1. Short-term false memories in recall tasks

Using a DRM-like paradigm with only four memory words and a retention interval of a few seconds, we reported relatively high rates of semantic intrusions in the immediate test in all experiments, averaging about 11%. This rate is slightly lower than the false alarm rate for semantic lures in recognition tasks (20% on average across studies), which is a finding that has also been reported in the LTM literature (e.g., Oliver et al., 2016; Seamon et al., 2002). Semantic errors were by far the most

frequent in the present study, with only between 1.2% and 2.4% of each type of non-semantic errors on average in the immediate tests. Atkins and Reuter-Lorenz (2008, Exp.1B) also reported a predominance of semantic errors over all other error types in a short-term DRM task with a recall test. This prevalence of semantic errors relative to other types of errors in a WM task with a recall test mirrors the results repeatedly obtained with recognition tests (e.g., Abadie & Camos, 2019; Atkins & Reuter-Lorenz, 2008; Flegal et al., 2010, 2014). This underscores the fact that short-term false memories do not depend solely on retrieval processes that are specific to recognition tasks.

By examining the confidence or subjective experience associated with each response, Experiments 3 and 4 provided an insight into the processes underlying these short-term false memories. In Experiment 3, where participants reported their confidence in their responses, both semantic and non-semantic errors in the immediate test were predominantly associated with low confidence, suggesting that confidence judgments may not be sensitive enough to distinguish the nature of short-term memory errors. In contrast, Experiment 4, in which participants reported their subjective experience rather than their confidence in their response, showed that short-term semantic errors were more often associated with a sense of recollecting the details of the presentation of the falsely recalled words or with a vague memory of the meaning of the falsely recalled words than other types of memory errors. Flegal et al. (2010, Exp. 2) also showed that false recognition of related items was more often associated with "remember" judgments (31%) and with "know" judgments (31%) than false recognition of unrelated items, which was more often associated with "guess" judgments (50%). Thus, assessing participants' subjective experience seems to reveal interesting differences in the representations underlying the two types of errors. These findings are consistent with those of Abadie and Camos (2019); Abadie & Rousselle, 2023 showing that retrieval of gist representations underlies short-term false recognition of related items and suggest that the processes underlying the occurrence of semantic errors in recall tests appear to be similar to those underlying them in recognition tasks.

Comparing the errors produced in the immediate test and those produced in the delayed test, there was an increase in errors in the delayed test due to the forced recall procedure in Experiments 2 to 4. Experiments 3 and 4 revealed that the increase in semantic errors was associated with an increase in confidence or in the feeling of retrieving details from the study phase, whereas non-semantic errors were more often associated with guess judgments. These results are consistent with findings in the LTM literature showing that false memories produced in an initial test can strengthen over time and multiple retrievals (e.g., Gallo, 2006, for a review). However, a direct comparison between immediate and delayed tests in our experiments should be viewed with caution because the recall tests differ not only in delay, but also in the fact that initial tests involved single lists, whereas the final test involved all lists.

Taken together, these results suggest that there are no major differences between short-term false memories obtained in recognition tasks and those obtained in recall tasks. Short-term false memories therefore do not seem to be determined by processes that take place during retrieval, but rather by processes that would take place during information maintenance.

6.2. Articulatory rehearsal prevents short-term false memories

According to Abadie and Camos (2019) model of short- and long-term false memories, because rehearsal reinforces verbatim memories that do not persist much over time, its use should increase short-term correct recall and reduce the occurrence of short-term false memories but have no effect on long-term recall (Fig. 1). These predictions are fully supported by the results of the present study (Table 1). Consistently across experiments, reducing rehearsal opportunities reduced immediate correct recall (about 12% loss on average in Experiments 1 and 2). It had no effect on delayed recall, except on the conditionalized correct recall score, which decreased with more rehearsal opportunities. This latter effect is presumably due to the fact that rehearsal emphasizes surface features of memory items that are not maintained over the long term. More importantly, as expected, limiting the use of rehearsal substantially increased semantic intrusions in the immediate test (about 7.5% more on average), while having virtually no effect on the other types of errors (<1% on average). Finally, this mechanism had no effect on the errors produced in the delayed test. However, when considering errors generated only during this test (i.e., E_{cond}), all types of errors, both semantic and non-semantic, increased with more rehearsal opportunities, which is the likely corollary of its effect on conditionalized correct recall. Results of previous studies using recognition tasks (e.g., Abadie & Camos, 2019; Atkins et al., 2011; Macé & Caza, 2011) echoed nicely with the present findings in recall tasks in which errors were spontaneously generated. Thus, active WM maintenance by means of rehearsal moderates the occurrence of short-term false memories, regardless of the retrieval task that is used.

The effects of manipulating rehearsal are also indicative of the recall processes on which rehearsal can act. The reduction in the incidence of semantic errors combined with the increase in correct recall in the immediate test when there were more rehearsal opportunities is consistent with the hypothesis that rehearsal emphasizes direct access to verbatim representations of memory items. Other studies also support this hypothesis by showing that rehearsal reinforces the surface, either phonological or articulatory, features of the words to be remembered (Baddeley, 1966; Camos et al., 2011, 2013; Estes, 1973) and reduces the importance of semantic processing (e.g., Higgins & Johnson, 2013; Loaiza & Camos, 2018; Oberauer, 2009). Finally, rehearsal had only transient and not a long-term beneficial effect in the present study, which is also consistent with the hypothesis that rehearsal favors the maintenance of item surface features that fade faster than semantic representations over time (Seamon et al., 2002).

6.3. Attentional refreshing fosters long-term false memories

According to Abadie and Camos (2019) model, the use of refreshing fosters the creation of both verbatim and gist traces, which should increase correct recall of studied items in immediate tests as well as correct recall of studied items and false recall of semantically related distractors in delayed tests. Consistently, across experiments, correct recall performance in the immediate test was better when there were more opportunities for refreshing. In the delayed test, although there was descriptively a greater increase of semantic errors when there were more refreshing opportunities, varying refreshing opportunities did not have a substantial effect on these errors. However, errors occurring in the delayed test can be errors that were first generated in the immediate test and then reproduced in the delayed test, as well as errors generated only in the delayed test. Interestingly, when considering errors generated only in the delayed test (i.e., E_{cond}), we found an increase in semantic errors in all four experiments when there were more refreshing opportunities (about 13% on average). Moreover, there was substantial evidence against the effect of refreshing on non-semantic errors in the delayed test in all experiments. Thus, it appears that the effect of refreshing on semantic errors in the delayed recall test is specifically visible for errors made only in this test. Consistently, Abadie and Camos (2019) found the same effect of refreshing on false recognition of related distractors in a delayed test in which the word lists tested were not previously tested in an immediate test.

A first question that comes to mind at this point is why refreshing increases the semantic errors produced specifically in the delayed test, whereas it seems to have no impact (or rather reduces them) in the immediate test. Some authors (e.g., Oberauer, 2013; Oberauer & Hein, 2012) have proposed that refreshing not only enhances memory for item-specific information, but also for item-context associations (e.g., the item and its serial position in the list for WM tasks with a serial recall test). Contextual cueing would allow memory items to be refreshed,

which would strengthen their accessibility in memory. Transposed to the DRM paradigm, the context would be the theme of a list (e.g., "mountain") and the items the words of the list (e.g., "hill, valley, climb, summit"). In such a view, without a cue as in recall tests, refreshing could function as a redintegration (Barrouillet & Camos, 2015; Hulme et al., 1997) or a reconstructive process (Brainerd et al., 2009; 2014; 2015; Brainerd & Reyna, 2010) using LTM knowledge. The reconstructive process would use partial identifying information (e.g., semantic features of items) to regenerate candidate items that would contain true candidates (e.g., "hill" or "valley") as well as false ones (e.g., "mountain"). This theoretical view is in line with the hypothesis that refreshing could both increase the probability of correct recall and also semantic errors in immediate and delayed tests. However, this does not account for the fact that refreshing did not increase semantic errors in the immediate test in our four experiments. An alternative view, consistent with the Abadie and Camos (2019) model, would be that refreshing enhances both direct access to item surface details (i.e., verbatim traces) and reconstructive retrieval processes (Fig. 1). In immediate tests, the surface details of the items are readily available in WM (as when rehearsal is used), which reduces the likelihood of semantic errors. In delayed tests, however, things are different for two reasons. First, the surface traces of studied items fade more rapidly than their semantic traces, leading to a net increase in semantic intrusions on delayed tests (Brainerd & Reyna, 2005). Second, the refreshing of items in WM is likely to leave partial traces of items in LTM (Barrouillet & Camos, 2015); these partial traces would reflect the semantic content of the items and thus promote semantic intrusions in delayed tests.

A second question is why the effect of refreshing on long-term semantic errors appeared only on those produced exclusively in the delayed test. A plausible response is that its effect is masked because of the influence of the prior immediate test on the delayed test. Indeed, as discussed above, it is well established that prior retrieval influences later learning (e.g., Chan, Manley, et al., 2018; Chan, Meissner, & Davis, 2018). The influence of prior recall on semantic intrusions might be stronger than the influence of the manipulation of refreshing opportunities.

Finally, one question remains as to why, in our studies, refreshing did not improve correct recall in the delayed test even when controlling for test-retest. Previous studies showed a decrease in correct recall in delayed tests when reducing refreshing opportunities (e.g., Camos & Portrat, 2015; Loaiza & McCabe, 2012, 2013; McCabe, 2008). However, some authors have argued that the improvement in delayed recall performance could be due to processes that take place during item encoding or to other processes such as elaboration, which would be different from refreshing (Bartsch et al., 2018; Loaiza et al., 2023; Souza & Oberauer, 2017). A difference between these studies and ours is the use of DRM lists. This type of lists could favor semantic context retrieval at the expense of item memory. Similarly, Abadie and Camos (2019) also found that only semantic false recognitions were impacted by the use of refreshing in delayed tests. Further studies are needed to better understand the role of refreshing on delayed recall.

6.4. Dissociation or overlap between WM and LTM processes?

Previous studies demonstrate striking parallel in the memory distortions that affect recognition over the short and long term (e.g., Atkins et al., 2011; Flegal et al., 2010). This led the authors to propose that the false memories observed in WM tasks arise from processes shared by both WM and LTM, bringing support to the theoretical view of a strong overlap between the two memory systems. However, some of the current findings were at odds with this theoretical view. The semantic intrusions that were obtained in the immediate test could not be generated by WM processes. In all four experiments, semantic errors were far more frequent when WM maintenance was impeded (about 15%) than when it was not (about 4.5%). This finding replicates the one obtained previously with recognition tasks (Abadie & Camos, 2019) and generalizes it by showing that it did not depend on the memory test. It seems, therefore, that the occurrence of false memories in short-term tests results from the weakness or absence of memory traces stored in WM when WM maintenance mechanisms cannot be used effectively to maintain them. Therefore, short-term false memories would merely reflect the influence of LTM, i.e., a store from which traces can be searched and retrieved in the absence of WM memory traces (e.g., Unsworth & Engle, 2007). These findings are thus difficult to reconcile with a unitary view of memory.

Alternatively, some models conceive WM as the activated part of LTM (Cowan, 1999, Cowan et al., 2021; Engle et al., 2001; Oberauer, 2002). For example, Cowan's (1999) embedded-processes model proposes that WM is a subset of LTM in which a set of elements relevant for the current task are activated among which three to four chunks of information are maintained in a high state of accessibility by a limited capacity focus of attention. When DRM lists are studied, the theme has a high probability of being activated (i.e., in the activated part of the LTM, but outside the focus of attention), due to the network strongly activated by the words studied (i.e., maintained in the focus of attention) all associated with this common theme (Stadler et al., 1999). Therefore, false memories based on the retrieval of memory traces in the activated part of LTM should be enhanced when refreshing opportunities are reduced (i.e., when maintenance of memory traces in the focus of attention is impaired). However, the present findings contradict this prediction. Varying the opportunities for refreshing had no effect on semantic errors in the immediate test, and semantic errors in the delayed test increased with more opportunities for refreshing.

Rather, the findings of the present study seem to support a conception of memory that can be described as dual (e.g., Norris, 2017), borrowing notions from earlier models of WM but distinguishing itself by integrating more closely with current neurocognitive models of consciousness and LTM. According to this view, which is consistent but goes beyond the first proposal by Abadie and Camos (2019), memory is conceived as a system in which attention is the main general resource for maintaining both short- and long-term information. The use of attention allows items to enter WM and become conscious, in the sense that they can be used flexibly by other processes (e.g., Baars, 1997; Dehaene et al., 2011). However, verbal information might also be maintained in the short term by a mechanism or system dissociated from this memory system, the articulatory loop. As proposed in the latest version of the TBRS model, the articulatory loop is conceived as a non-attentional mechanism dedicated to the maintenance of motor programs through rehearsal that allows the reproduction of verbal items through articulation (Barrouillet et al., 2021; Barrouillet & Camos, 2021). According to this view, the use of rehearsal to maintain semantically related information in the paradigm used in the present study would prevent the occurrence of semantic memory errors by providing direct access to the surface, articulatory form of each word. These errors would occur in the short term, especially when rehearsal is not or cannot be used (Abadie & Camos, 2019). Other general attentional mechanisms, such as but not limited to refreshing, can be used to maintain verbal information independently of the articulatory loop (e.g., Camos, 2015; Camos, 2017). As discussed above, these mechanisms enable the maintenance of verbatim and gist representations that are differentially affected by time as proposed in the FTT (e.g., Brainerd et al., 2015). Although it accounts for the results obtained in this study, this conception briefly sketched here needs to be developed and tested in future studies.

6.5. Limitations

The present study, however, has some limitations that may affect the generalizability of the findings. The first is that, as in many psychological studies, the participants were mainly university students with a level of education that may be higher than the average level of the general population. The second limitation is the gender imbalance, with women predominant in all experiments. This lack of gender diversity could

influence the results obtained, although other studies have found no effect of gender on the occurrence of false memories from neutral word lists (Dewhurst et al., 2012). Finally, interindividual differences in WM capacity were not considered in the present study. The WM task may have been more or less attentionally demanding depending on the individual, which could, for example, make the refreshing manipulation less effective for individuals with greater WM capacity. It should be noted, however, that the refreshing manipulation still proved effective for immediate correct recall in all experiments. To overcome this limitation, future studies could use a titration procedure to adapt the task to individual WM capacities.

7. Concluding comments

Accumulated evidence supports the conclusion that the semantic DRM illusion can occur in WM tasks with lists of only a few items and an interval of a few seconds between study and test. The present study investigated the role of WM maintenance mechanisms in this phenomenon, testing for the first time Abadie and Camos (2019) model in recall tasks. Importantly, this study showed that short-term semantic errors are more frequent when WM maintenance is impeded, i.e., when recall relies primarily on LTM retrieval. Specifically, articulatory rehearsal prevents short-term semantic errors. In contrast, attentional refreshing has no specific effect on short-term semantic errors, but favors the occurrence of semantic errors that appear only in the long term. These findings are consistent with a dual conception of memory, consisting of a central attentional system for the maintenance of short-and long-term information, and an independent system based on articulatory rehearsal for the maintenance of short-term verbal information. Although gist representations can be maintained in WM, they are not emblematic of WM in which direct access processes that retrieve surface memory traces appears to be used preferentially.

Funding source

This work was supported by a grant from the French National Research Agency (No. ANR-19-28–0017CE-01) to Marlène Abadie. The funding source was not involved in the conduct of the research nor in the preparation of the article.

CRediT authorship contribution statement

Marlène Abadie: Writing – review & editing, Writing – original draft, Visualization, Supervision, Software, Project administration, Methodology, Funding acquisition, Formal analysis, Data curation, Conceptualization. **Christelle Guette:** Software, Investigation, Data curation. **Amélie Troubat:** Investigation, Data curation. **Valérie Camos:** Writing – review & editing, Methodology, Conceptualization.

Data availability

Materials, data, and additional analyses are available on OSF at https://osf.io/rqkvw/?view_only=6dad2a1829db458c9b68bd36f614ddd3

References

Abadie, M., & Camos, V. (2019). False memory at short and long term. *Journal of Experimental Psychology: General, 148*(8), 1312–1334. https://doi.org/10.1037/xge0000526

Abadie, M., & Rousselle, M. (2023). Short-term phantom recollection in 8–10-year-olds and young adults. *Journal of Intelligence, 11*(4), 67. https://doi.org/10.3390/jintelligence11040067

Abadie, M., Waroquier, L., & Terrier, P. (2013). Gist memory in the unconscious thought effect. *Psychological Science, 24*, 1253–1259. https://doi.org/10.1177/0956797612470958

Abadie, M., Waroquier, L., & Terrier, P. (2017). The role of gist and verbatim memory in complex decision making: Explaining the unconscious-thought effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 43*, 694–705. https://doi.org/10.1037/xlm0000336

Allen, R. J., Hitch, G. J., & Baddeley, A. D. (2018). Exploring the sentence advantage in working memory: Insights from serial recall and recognition. *Quarterly Journal of Experimental Psychology, 71*(12), 2571–2585. https://doi.org/10.1177/1747021817746929

Atkins, A. S., Berman, M. G., Reuter-Lorenz, P. A., Lewis, R. L., & Jonides, J. (2011). Resolving semantic and proactive interference in memory over the short-term. *Memory & Cognition, 39*(5), 806–817. https://doi.org/10.3758/s13421-011-0072-5

Atkins, A. S., & Reuter-Lorenz, P. A. (2008). False working memories? Semantic distortion in a mere 4 seconds. *Memory & Cognition, 36*(1), 74–81. https://doi.org/10.3758/MC.36.1.74

Atkins, A. S., & Reuter-Lorenz, P. A. (2011). Neural mechanisms of semantic interference and false recognition in short-term memory. *NeuroImage, 56*(3), 1726–1734. https://doi.org/10.1016/j.neuroimage.2011.02.048

Baars, B. J. (1997). In the theatre of consciousness. Global workspace theory, a rigorous scientific theory of consciousness. *Journal of Consciousness Studies, 4*(4), 292–309.

Baddeley, A. (2012). Working memory: Theories, models, and controversies. *Annual Review of Psychology, 63*, 1–29. https://doi.org/10.1146/annurev-psych-120710-100422

Baddeley, A. D. (1986). *Working memory*. Oxford: Clarendon Press.

Baddeley, A. D. (2007). *Working memory, thought, and action*. Oxford: Oxford University Press.

Barrouillet, P., Bernardin, S., & Camos, V. (2004). Time constraints and resource sharing in adults' working memory spans. *Journal of Experimental Psychology: General, 133*(1), 83–100. https://doi.org/10.1037/0096-3445.133.1.83

Barrouillet, P., Bernardin, S., Portrat, S., Vergauwe, E., & Camos, V. (2007). Time and cognitive load in working memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 33*(3), 570–585. https://doi.org/10.1037/0278-7393.33.3.570

Barrouillet, P., & Camos, V. (2015). *Working memory: Loss and reconstruction*. Hove, UK: Psychology Press.

Barrouillet, P., & Camos, V. (2021). The time-based resource-sharing model of working memory. In R. H. Logie, V. Camos, & N. Cowan (Eds.), *Working memory: State of the science* (pp. 85–115). Oxford University Press. https://doi.org/10.1093/oso/9780198842286.003.0004.

Barrouillet, P., Gorin, S., & Camos, V. (2021). Simple spans underestimate verbal working memory capacity. *Journal of Experimental Psychology: General, 150*(4), 633–665. https://doi.org/10.1037/xge0000957

Bartsch, L. M., Singmann, H., & Oberauer, K. (2018). The effects of refreshing and elaboration on working memory performance, and their contributions to long-term memory formation. *Memory & Cognition, 46*(5), 796–808. https://doi.org/10.3758/s13421-018-0805-9

Bonin, P., Méot, A., Ferrand, L., & Bugaïska, A. (2013). Normes d'associations verbales pour 520 mots concrets et étude de leurs relations avec d'autres variables psycholinguistiques. *L'Année Psychologique, 113*, 63–92. https://doi.org/10.3917/anpsy.131.0063

Brainerd, C. J., Aydin, C., & Reyna, V. F. (2012). Development of dual-retrieval processes in recall: Learning, forgetting, and reminiscence. *Journal of Memory and Language, 66*(4), 763–788. https://doi.org/10.1016/j.jml.2011.12.002

Brainerd, C. J., Gomes, C. F. A., & Moran, R. (2014). The two recollections. *Psychological Review, 121*(4), 563–599. https://doi.org/10.1037/a0037668

Brainerd, C. J., Gomes, C. F. A., & Nakamura, K. (2015). Dual recollection in episodic memory. *Journal of Experimental Psychology: General, 144*(4), 816–843. https://doi.org/10.1037/xge0000084

Brainerd, C. J., & Reyna, V. F. (2005). *The science of false memory*. Oxford University Press.

Brainerd, C. J., & Reyna, V. F. (2010). Recollective and nonrecollective recall. *Journal of Memory and Language, 63*(3), 425–445. https://doi.org/10.1016/j.jml.2010.05.002

Brainerd, C. J., Reyna, V. F., & Howe, M. L. (2009). Trichotomous processes in early memory development, aging, and neurocognitive impairment: A unified theory. *Psychological Review, 116*(4), 783–832. https://doi.org/10.1037/a0016963

Brainerd, C. J., Wright, R., Reyna, V. F., & Mojardin, A. H. (2001). Conjoint recognition and phantom recollection. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 27*(2), 307–327. https://doi.org/10.1037/0278-7393.27.2.307

Brainerd, C. J., Yang, Y., Reyna, V. F., Howe, M. L., & Mills, B. A. (2008). Semantic processing in "associative" false memory. *Psychonomic Bulletin & Review, 15*(6), 1035–1053. https://doi.org/10.3758/PBR.15.6.1035

Camos, V. (2015). Storing verbal information in working memory. *Current Directions in Psychological Science, 24*(6), 440–445. https://doi.org/10.1177/0963721415606630

Camos, V. (2017). Domain-specific versus domain-general maintenance in working memory. In B. Ross (Ed.), *Vol. 67. The psychology of learning and motivation* (pp. 135–171). Cambridge, MA: Academic Press. https://doi.org/10.1016/bs.plm.2017.03.005.

Camos, V., & Barrouillet, P. (2014). Attentional and non-attentional systems in the maintenance of verbal information in working memory: The executive and phonological loops. *Frontiers in Human Neuroscience, 8*, 900. https://doi.org/10.3389/fnhum.2014.00900

Camos, V., Johnson, M., Loaiza, V., Portrat, S., Souza, A., & Vergauwe, E. (2018). What is attentional refreshing in working memory? *Annals of the New York Academy of Sciences, 1424*(1), 19–32. https://doi.org/10.1111/nyas.13616

Camos, V., Lagner, P., & Barrouillet, P. (2009). Two maintenance mechanisms of verbal information in working memory. *Journal of Memory and Language, 61*(3), 457–469. https://doi.org/10.1016/j.jml.2009.06.002

Camos, V., Mora, G., & Oberauer, K. (2011). Adaptive choice between articulatory rehearsal and attentional refreshing in verbal working memory. *Memory & Cognition, 39*, 231–244. https://doi.org/10.3758/s13421-010-0011-x

Camos, V., & Portrat, S. (2015). The impact of cognitive load on delayed recall. *Psychonomic Bulletin & Review, 22*, 1029–1034. https://doi.org/10.3758/s13423-014-0772-5

Chan, J. C., Manley, K. D., Davis, S. D., & Szpunar, K. K. (2018). Testing potentiates new learning across a retention interval and a lag: A strategy change perspective. *Journal of Memory and Language, 102*, 83–96. https://doi.org/10.1016/j.jml.2018.05.007

Chan, J. C., Meissner, C. A., & Davis, S. D. (2018). Retrieval potentiates new learning: A theoretical and meta-analytic review. *Psychological Bulletin, 144*, 1111–1146. https://doi.org/10.1037/bul0000166

Chang, M., & Brainerd, C. (2021). Semantic and phonological false memory: A review of theory and data. *Journal of Memory and Language, 119*, Article 104210. https://doi.org/10.1016/j.jml.2020.104210

Chen, Z., & Cowan, N. (2009). How verbal memory loads consume attention. *Memory & Cognition, 37*, 829–836. https://doi.org/10.3758/MC.37.6.829

Coane, J. H., McBride, D. M., Raulerson, B. A., III, & Jordan, J. S. (2007). False memory in a short-term memory task. *Experimental Psychology, 54*, 62–70. https://doi.org/10.1027/1618-3169.54.1.62

Cowan, N. (1999). An embedded-processes model of working memory. In A. Miyake, & P. Shah (Eds.), *Models of working memory: Mechanisms of active maintenance and executive control* (pp. 62–101). Cambridge University Press.

Cowan, N., Morey, C. C., & Naveh-Benjamin, M. (2021). An embedded-processes approach to working memory: How is it distinct from other approaches, and to what ends? In R. H. Logie, V. Camos, & N. Cowan (Eds.), *Working memory: State of the science* (pp. 44–84). Oxford University Press.

Craik, F. I. M., & Lockhart, R. S. (1972). Levels of processing: A framework for memory research. *Journal of Verbal Learning and Verbal Behavior, 11*, 671–684. https://doi.org/10.1016/S0022-5371(72)80001-X

Dehaene, S., Changeux, J. P., & Naccache, L. (2011). The global neuronal workspace model of conscious access: From neuronal architectures to clinical applications. In S. Dehaene, & Y. Christen (Eds.), *Characterizing consciousness: From cognition to the clinic?* (pp. 55–84). Research and Perspective in Neurosciences. Springer.

Dewhurst, S. A., Anderson, R. J., & Knott, L. M. (2012). A gender difference in the false recall of negative words: Women DRM more than men. *Cognition and Emotion, 26*, 65–74. https://doi.org/10.1080/02699931.2011.553037

Engle, R. W. (2001). What is working memory capacity? In H. L. Roediger, III, J. S. Nairne, I. Neath, & A. M. Surprenant (Eds.), *The nature of remembering: Essays in honor of Robert G. Crowder (pp. 297–314)*. American Psychological Association.

Estes, W. K. (1973). Phonemic coding and rehearsal in short-term memory for letter strings. *Journal of Verbal Learning and Verbal Behavior, 12*(4), 360–372. https://doi.org/10.1016/S0022-5371(73)80015-5

Faul, F., Erdfelder, E., Lang, A. G., & Buchner, A. (2007). G* power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods, 39*(2), 175–191. https://doi.org/10.3758/BF03193146

Finger, H., Goeke, C., Diekamp, D., Standvoß, K., & König, P. (2017). LabVanced: A unified JavaScript framework for online studies. In *International conference on computational social science (Cologne)*.

Flegal, K. E., Atkins, A. S., & Reuter-Lorenz, P. A. (2010). False memories seconds later: The rapid and compelling onset of illusory recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 36*(5), 1331–1338. https://doi.org/10.1037/a0019903

Flegal, K. E., & Reuter-Lorenz, P. A. (2014). Get the gist? The effects of processing depth on false recognition in short-term and long-term memory. *Memory & Cognition, 42*(5), 701–711. https://doi.org/10.3758/s13421-013-0391-9

Gallo, D. A. (2006). *Associative illusions of memory: Research on false memory for related events*. Hove, UK: Psychology Press.

Gallo, D. A. (2010). False memories and fantastic beliefs: 15 years of the DRM illusion. *Memory & Cognition, 38*(7), 833–848. https://doi.org/10.3758/MC.38.7.833

Greene, R. L. (1987). Effects of maintenance rehearsal on human memory. *Psychological Bulletin, 102*, 403–413. https://doi.org/10.1037/0033-2909.102.3.403

Higgins, J. A., & Johnson, M. K. (2013). Lost thoughts: Implicit semantic interference impairs reflective access to currently active information. *Journal of Experimental Psychology: General, 142*(1), 6–11. https://doi.org/10.1037/a0028191

Hulme, C., Roodenrys, S., Schweickert, R., Brown, G. D., Martin, S., & Stuart, G. (1997). Word-frequency effects on short-term memory tasks: Evidence for a redintegration process in immediate serial recall. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 23*(5), 1217–1232. https://doi.org/10.1037/0278-7393.23.5.1217

JASP Team. (2022). *JASP (Version 0.16.3) [Computer software]*.

Johnson, M. K., Raye, C. L., Mitchell, K. J., Greene, E. J., Cunningham, W. A., & Sanislow, C. A. (2005). Using fMRI to investigate a component process of reflection: Prefrontal correlates of refreshing a just activated representation. *Cognitive, Affective, & Behavioral Neuroscience, 5*, 39–361. https://doi.org/10.3758/CABN.5.3.339

Jones, D., Farrand, P., Stuart, G., & Morris, N. (1995). Functional equivalence of verbal and spatial information in serial short-term memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 21*(4), 1008–1018. https://doi.org/10.1037/0278-7393.21.4.1008

Jones, D. M., Hughes, R. W., & Macken, W. J. (2007). The phonological store abandoned. *Quarterly Journal of Experimental Psychology, 60*(4), 505–511. https://doi.org/10.1080/17470210601147591

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association, 90*, 773–795. https://doi.org/10.1080/01621459.1995.10476572

Kruschke, J. K. (2011). Bayesian assessment of null values via parameter estimation and model comparison. *Perspectives on Psychological Science, 6*(3), 299–312. https://doi.org/10.1177/1745691611406925

Kruschke, J. K., & Liddell, T. M. (2018). The Bayesian new statistics: Hypothesis testing, estimation, meta-analysis, and power analysis from a Bayesian perspective. *Psychonomic Bulletin & Review, 25*, 178–206. https://doi.org/10.3758/s13423-016-1221-4

Lampinen, J. M., Neuschatz, J. S., & Payne, D. G. (1998). Memory illusions and consciousness: Examining the phenomenology of true and false memories. *Current Psychology, 16*, 181–223. https://doi.org/10.1007/s12144-997-1000-5

Lampinen, J. M., Ryals, B. D., & Smith, K. (2008). Compelling untruths: The effect of retention interval on content borrowing and vivid false memories. *Memory, 16*(2), 149–156. https://doi.org/10.1080/09658210701839277

Loaiza, V. M., & Camos, V. (2018). The role of semantic representations in verbal working memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 44*, 863–881. https://doi.org/10.1037/xlm0000475

Loaiza, V. M., & McCabe, D. P. (2012). Temporal-contextual processing in working memory: Evidence from delayed cued recall and delayed free recall tests. *Memory & Cognition, 40*, 191–203. https://doi.org/10.3758/s13421-011-0148-2

Loaiza, V. M., & McCabe, D. P. (2013). The influence of aging on attentional refreshing and articulatory rehearsal during working memory on later episodic memory performance. *Aging, Neuropsychology, and Cognition, 20*(4), 471–493. https://doi.org/10.1080/13825585.2012.738289

Loaiza, V. M., Oftinger, A.-L., & Camos, V. (2023). How does working memory promote traces in episodic memory? *Journal of Cognition, 6*(1), 4. https://doi.org/10.5334/joc.245

Logie, R. H., Camos, V., & Cowan, N. (2021). Working memory: State of the science: An introduction. In R. H. Logie, V. Camos, & N. Cowan (Eds.), *Working memory: State of the science* (pp. 1–9). Oxford, UK: Oxford University Press.

Macé, A.-L., & Caza, N. (2011). The role of articulatory suppression in immediate false recognition. *Memory, 19*(8), 891–900. https://doi.org/10.1080/09658211.2011.613844

Macken, W. J., Taylor, J. C., Kozlov, M. D., Hughes, R. W., & Jones, D. M. (2016). Memory as embodiment: The case of modality and serial short-term memory. *Cognition, 155*, 113–124. https://doi.org/10.1016/j.cognition.2016.06.013

Malmberg, K. J. (2008). Recognition memory: A review of the critical findings and an integrated theory for relating them. *Cognitive Psychology, 57*(4), 335–384. https://doi.org/10.1016/j.cogpsych.2008.02.004

McCabe, D. P. (2008). The role of covert retrieval in working memory span tasks: Evidence from delayed recall tests. *Journal of Memory and Language, 58*, 480–494. https://doi.org/10.1016/j.jml.2007.04.004

Naveh-Benjamin, M., & Jonides, J. (1984). Maintenance rehearsal: A two-component analysis. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 10*, 369–385. https://doi.org/10.1037/0278-7393.10.3.369

Nelson, D. L., McEvoy, C. L., & Schreiber, T. A. (2004). The University of South Florida free association, rhyme, and word fragment norms. *Behavior Research Methods, Instruments, & Computers, 36*(3), 402–407. https://doi.org/10.3758/BF03195588

Norris, D. (2017). Short-term memory and long-term memory are still different. *Psychological Bulletin, 143*(9), 992–1009. https://doi.org/10.1037/bul0000108

Oberauer, K. (2002). Access to information in working memory: Exploring the focus of attention. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 28*(3), 411–421. https://doi.org/10.1037/0278-7393.28.3.411

Oberauer, K. (2009). Interference between storage and processing in working memory: Feature overwriting, not similarity-based competition. *Memory & Cognition, 37*(3), 346–357. https://doi.org/10.3758/MC.37.3.346

Oberauer, K. (2013). The focus of attention in working memory—From metaphors to mechanisms. *Frontiers in Human Neuroscience, 7*, 673. https://doi.org/10.3389/fnhum.2013.00673

Oberauer, K., & Hein, L. (2012). Attention to information in working memory. *Current Directions in Psychological Science, 21*(3), 164–169. https://doi.org/10.1177/0963721412444727

Oliver, M. C., Bays, R. B., & Zabrucky, K. M. (2016). False memories and the DRM paradigm: Effects of imagery, list, and test type. *The Journal of General Psychology, 143*(1), 33–48. https://doi.org/10.1080/00221309.2015.1110558

Payne, D. G., Elie, C. J., Blackwell, J. M., & Neuschatz, J. S. (1996). Memory illusions: Recalling, recognising, and recollecting events that never occurred. *Journal of Memory & Language, 35*, 261–285. https://doi.org/10.1006/jmla.1996.0015

Raye, C. L., Johnson, M. K., Mitchell, K. J., Greene, E. J., & Johnson, M. R. (2007). Refreshing: A minimal executive function. *Cortex, 43*, 135–145. https://doi.org/10.1016/S0010-9452(08)70451-9

Reyna, V. F., & Brainerd, C. J. (1995). Fuzzy-trace theory: An interim synthesis. *Learning and Individual Differences, 7*, 1–75. https://doi.org/10.1016/1041-6080(95)90031-4

Roediger, H. L., & McDermott, K. B. (1995). Creating false memories: Remembering words not presented in lists. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 21*(4), 803–814. https://doi.org/10.1037/0278-7393.21.4.803

Roediger, H. L., & Payne, D. G. (1985). Recall criterion does not affect recall level or hypermnesia: A puzzle for generate/recognize theories. *Memory & Cognition, 13*, 1–7. https://doi.org/10.3758/BF03198437

Roediger, H. L., III, Wheeler, M. A., & Rajaram, S. (1993). Remembering, knowing, and reconstructing the past. In *, Vol. 30. Psychology of learning and motivation* (pp. 97–134). Academic Press. https://doi.org/10.1016/S0079-7421(08)60295-9.

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

Rousselle, M., Abadie, M., Blaye, A., & Camos, V. (2022). Children's gist-based false memory in working memory tasks. *Developmental Psychology, 59*, 272–284. https://doi.org/10.1037/dev0001476

Seamon, J. G., Luo, C. R., Kopecky, J. J., Price, C. A., Rothschild, L., Fung, N. S., & Schwartz, M. A. (2002). Are false memories more difficult to forget than accurate memories?: The effect of retention interval on recall and recognition. *Memory & Cognition, 30*(7), 1054–1064. https://doi.org/10.3758/bf03194323

Smith, E. E., & Jonides, J. (1999). Storage and executive processes in the frontal lobes. *Science, 283*, 1657–1661. https://doi.org/10.1126/science.283.5408.1657

Souza, A. S., & Oberauer, K. (2017). Time to process information in working memory improves episodic memory. *Journal of Memory and Language, 96*, 155–167. https://doi.org/10.1016/j.jml.2017.07.002

Stadler, M. A., Roediger, H. L., & McDermott, K. B. (1999). Norms for word lists that create false memories. *Memory & Cognition, 27*(3), 494–500. https://doi.org/10.3758/BF03211543

Trost, S., & Gruber, O. (2012). Evidence for a double dissociation of articulatory rehearsal and non-articulatory maintenance of phonological information in human verbal working memory. *Neuropsychobiology, 65*, 133–140. https://doi.org/10.1159/000332335

Tulving, E. (1985). Memory and consciousness. *Canadian Psychology / Psychologie Canadienne, 26*(1), 1–12. https://doi.org/10.1037/h0080017

Uittenhove, K., Chaabi, L., Camos, V., & Barrouillet, P. (2019). Is working memory storage intrinsically domain-specific? *Journal of Experimental Psychology: General, 148*(11), 2027–2057. https://doi.org/10.1037/xge0000566

Unsworth, N., & Engle, R. W. (2007). The nature of individual differences in working memory capacity: Active maintenance in primary memory and controlled search from secondary memory. *Psychological Review, 114*(1), 104–132. https://doi.org/10.1037/0033-295X.114.1.104

Wixted, J. T. (2007). Dual-process theory and signal-detection theory of recognition memory. *Psychological Review, 114*(1), 152–176. https://doi.org/10.1037/0033-295X.114.1.152

Yonelinas, A. P. (2002). The nature of recollection and familiarity: A review of 30 years of research. *Journal of Memory and Language, 46*(3), 441–517. https://doi.org/10.1006/jmla.2002.2864