

**The Role of Consolidation in Interference-Based Forgetting:
A Critical Re-Evaluation of Behavioral Evidence**

Julian Quevedo Pütter

Inaugural Dissertation

Submitted in partial fulfillment of the requirements for the degree of Doctor of Social Sciences in the DFG Research Training Group “Statistical Modeling in Psychology”
at the University of Mannheim

Main Supervisor:

Prof. Dr. Edgar Erdfelder

Additional Supervisors:

Prof. Dr. Mandy Hütter

Prof. Dr. Beatrice G. Kuhlmann

Dean of the School of Social Sciences:

Prof. Dr. Michael Diehl

Thesis Evaluators:

Prof. Dr. Arndt Bröder

Prof. Dr. Beatrice G. Kuhlmann

Thesis Defense:

October 15, 2024

Contents

Summary	V
Manuscripts	VII
1 Introduction	1
1.1 Interference-Based Accounts of Forgetting	2
1.2 Empirical Findings on Interference-Based Forgetting	6
2 Storage-Retrieval Analyses	11
2.1 Multinomial Processing Tree Modeling	12
2.2 Sleep-Induced Retrograde Facilitation	15
3 Re-Evaluation of Behavioral Evidence	19
3.1 Alcohol-Induced Retrograde Facilitation	19
3.2 Temporal Gradient of Retroactive Interference	23
3.3 Rest-Induced Retrograde Facilitation	28
4 Discussion	33
4.1 Synthesis	33
4.2 A Revised Model of Interference-Based Forgetting	40
4.3 Strengths and Limitations	43
4.4 Future Research	46
4.5 Conclusion	48
5 Bibliography	49
A Acknowledgements	57
B Statement of Originality	59
C Co-Authors' Statements	61
D Copies of Manuscripts	65

Summary

The inhibition of memory consolidation has been proposed to explain a wide range of interference-based forgetting phenomena (Wixted, 2004). According to the opportunistic consolidation account (Mednick et al., 2011), reducing, delaying, or minimizing retroactive interference provides ideal conditions for consolidation processes to unfold. At the same time, passive interference reduction accounts that dispense with a consolidation mechanism have been largely neglected. According to the temporal distinctiveness account (Brown et al., 2007), reducing, delaying, or minimizing retroactive interference increases the isolation of memory representations on the temporal dimension of a psychological memory space. Crucially, whereas the opportunistic consolidation account attributes interference-based forgetting to storage processes, the temporal distinctiveness account explains forgetting in terms of retrieval processes.

In this thesis, I adopt a storage-retrieval multinomial processing tree (MPT) modeling approach to precisely disentangle storage and retrieval contributions to memory performance following reduced, delayed, or minimized retroactive interference. In the first manuscript (Quevedo Pütter & Erdfelder, 2022), we report an experiment that investigated alcohol-induced retrograde facilitation. Reduced retroactive interference in an alcohol compared to a placebo condition resulted in significantly higher retrieval but not storage probabilities for the previously encoded word pairs. In the second manuscript (Quevedo Pütter et al., 2024), we scrutinized the mechanisms underlying the temporal gradient of retroactive interference. In three experiments, participants engaged in interpolated learning either relatively early or relatively late during the retention interval. Delaying retroactive interference again resulted in increased retrieval but not storage probabilities. Finally, in the third manuscript (Quevedo Pütter & Erdfelder, 2024), we intended to effectively minimize retroactive interference by means of post-encoding waking rest. In two experiments, participants wakefully rested, used social media, or engaged in unrelated vocabulary learning after the original learning phase. In contrast to the first and second manuscript, we found rest-induced retrograde facilitation to be driven by storage processes.

Overall, this mixed result pattern indicates that interference-based forgetting can largely be explained in terms of retrieval processes. Opportunistic consolidation seems to be inhibited only under rather specific conditions. In light of these conclusions, I propose an integrative account of interference-based forgetting that combines elements of the temporal distinctiveness and the opportunistic consolidation account.

Manuscripts

This cumulative thesis is based on three manuscripts, one of which has been published and two of which have been submitted for publication. In addition, I refer to an unpublished manuscript that I have contributed to as a co-author.

Manuscript I

Quevedo Pütter, J., & Erdfelder, E. (2022). Alcohol-induced retrograde facilitation? Mixed evidence in a preregistered replication and encoding-maintenance-retrieval analysis. *Experimental Psychology*, *69*(6), 335-350. <https://doi.org/10.1027/1618-3169/a000569>

Manuscript II

Quevedo Pütter, J., Dahler, S., & Erdfelder, E. (2024). *Opportunistic consolidation or temporal distinctiveness? Retrieval, not storage, drives the temporal gradient of retroactive interference in episodic memory.* Manuscript submitted for publication.

Manuscript III

Quevedo Pütter, J., & Erdfelder, E. (2024). *Waking rest during retention facilitates memory consolidation, but so does social media use: A storage-retrieval analysis.* Manuscript submitted for publication.

Additional Manuscript

Erdfelder, E., Berres, S., Quevedo Pütter, J., & Küpper-Tetzl, C. E. (2022). *Why does sleep improve episodic memory? An encoding-maintenance-retrieval analysis.* Manuscript under revision.

“Such a cruel thing, memory. We can’t remember what it is that we’ve forgotten.”

—Margaret Atwood, *The Testaments*

1 Introduction

One of the few certainties about human episodic memory is its imperfection. Over a century ago, Ebbinghaus (1885) pioneered experimental memory research by systematically evaluating the rate of forgetting across different time intervals of up to 31 days after encoding. In accordance with subjective experience (Wixted, 2004) and later replications (e.g., White, 2001; Wixted & Ebbesen, 1991, 1997) Ebbinghaus' forgetting curve is characterized by a rapid decline in memory performance within the first minutes and hours, and a gradual flattening thereafter.

Although forgetting is ascribed various adaptive functions (e.g., emotion regulation, abstraction and automatization, context attunement), most people tend to be frustrated by the experience of not being able to remember some previously learned information (Nørby, 2015). Unsurprisingly, therefore, much psychological and neuroscientific research has followed up Ebbinghaus' (1885) groundbreaking work by investigating the reasons for *why* we forget what we once knew.

In a rather pessimistic interim conclusion, Wixted (2004) criticized that decades of memory research had ultimately not resulted in a coherent theory of forgetting, but had rather assembled an “atheoretical laundry list of factors that may or may not play a role” (p. 236). As a consequence, he put forward the tenets of a comprehensive interference-based account of forgetting that continues to inform memory research to this day. Essentially, he advocated a dominant role for consolidation processes to explain not only sleep-related memory benefits (see Berres & Erdfelder, 2021), but indeed a wide range of interference-related memory phenomena even during wakefulness.

Wixted's (2004) seminal contribution has sparked fundamental debates as to the exact role of consolidation in interference-based forgetting. Opposing theoretical accounts have been developed that either include consolidation as a key factor (Dewar et al., 2007; Mednick et al., 2011) or propose alternative mechanisms instead (Brown et al., 2007; Ecker et al., 2015). In the following, I discuss these accounts of interference-based forgetting and give an overview of the relevant empirical findings. As will become clear, the available evidence is insufficient for a well-founded decision between the competing accounts. In the three manuscripts included in this thesis, I address this important gap in the literature by replicating and re-evaluating key empirical findings by means of appropriate storage-retrieval multinomial processing tree (MPT) models (Küpper-Tetzel & Erdfelder, 2012; Riefer & Batchelder, 1995).

1.1 Interference-Based Accounts of Forgetting

Episodic memory has been defined as a declarative memory system that allows human beings to remember past experiences associated with their temporal and spatial context (“mental time travel,” Tulving, 1972, 2002). Interference-based accounts of forgetting assume successful recollection of such episodic information to be impaired by additional information encoded before or after the to-be-remembered information. Whereas proactive interference occurs when new learning is impaired by previous learning, retroactive interference occurs when interpolated learning impairs the ability to subsequently remember older pieces of information.

Although some authors have at times argued that most forgetting can be attributed to proactive interference (Underwood, 1957), the notion of retroactive interference has been especially powerful in explaining a wide range of episodic memory phenomena. First and foremost, sleep-induced retrograde facilitation, that is, the positive effect of post-encoding sleep on subsequent memory performance has been interpreted by cognitive psychologists in terms of minimal retroactive interference during sleep (see Berres & Erdfelder, 2021 for a review and meta-analysis). Jenkins and Dallenbach (1924) were the first to empirically demonstrate the sleep benefit in episodic memory by having their participants spend retention intervals of different lengths either awake or asleep. The authors observed substantially better recall in the sleep than in the wake condition and attributed this effect to differences in retroactive interference between conditions.

Memory research throughout the 20th century has established two allegedly fundamental principles of retroactive interference in episodic memory. First, the higher the similarity between original and interpolated learning materials, the stronger the interference effect is generally expected to be (but see Antony et al., 2022). For example, McGeoch and McDonald (1931) observed free recall of previously learned adjectives to be most impaired by the learning of interpolated synonyms, followed by antonyms, unrelated adjectives, nonsense syllables, numbers, and unrelated reading. Second, rather than eliminating the original information from memory, retroactive interference is assumed to merely impair its retrieval from memory. In a classic study, Tulving and Psotka (1971) had their participants learn word lists, with each list containing six categories of four words. Unsurprisingly, increasing the number of interpolated word lists impaired final free recall for the original list, both in terms of the number of recalled words and the number of categories represented by at least one recalled word. Crucially, however, the ratio of recalled words per category was

not systematically affected by the number of interpolated lists. This suggests that retroactive interference impaired the retrieval of categories, but did not reduce the number of words available per category.

In light of these two principles, traditional accounts typically attribute interference-based forgetting to some kind of response competition, that is, original and interpolated pieces of information competing for retrieval from memory (Bower et al., 1994). This idea has mostly been applied to cue-overload procedures such as the A-B, A-C paired associates learning paradigm. In this paradigm, interpolated A-C item pairs are presented that consist of cue words (A) already presented during the original learning phase and new target words (C) that replace the originally presented target words (B). By pairing a single cue word with multiple target words, it is assumed that the original target (B) must compete for retrieval with the interpolated targets (see Antony et al., 2022).

Based on a careful review of the literature, Wixted (2004) formulated an innovative account of interference-based forgetting that dismisses much of the traditional retroactive interference literature. More specifically, he argued that conventional experimental procedures such as the A-B, A-C learning paradigm were ill-suited for obtaining a better understanding of everyday forgetting outside the laboratory. Instead, he took up some of the pioneering yet mostly forgotten work reported over a century earlier by Müller and Pilzecker (1900). According to Wixted's reading of the literature, real-life forgetting is for the most part caused by nonspecific retroactive interference from everyday mental exertion rather than specific retroactive interference from highly similar interpolated learning. He assumed such nonspecific retroactive interference not to interfere with the retrieval of previously encoded information, but rather with its consolidation, that is, the initial stabilization and subsequent redistribution of newly created memory traces (Dudai, 2004; McGaugh, 2000). Thus, forgetting is attributed to the inhibition of vital consolidation processes and, as a direct consequence, the degradation of the respective memory traces.

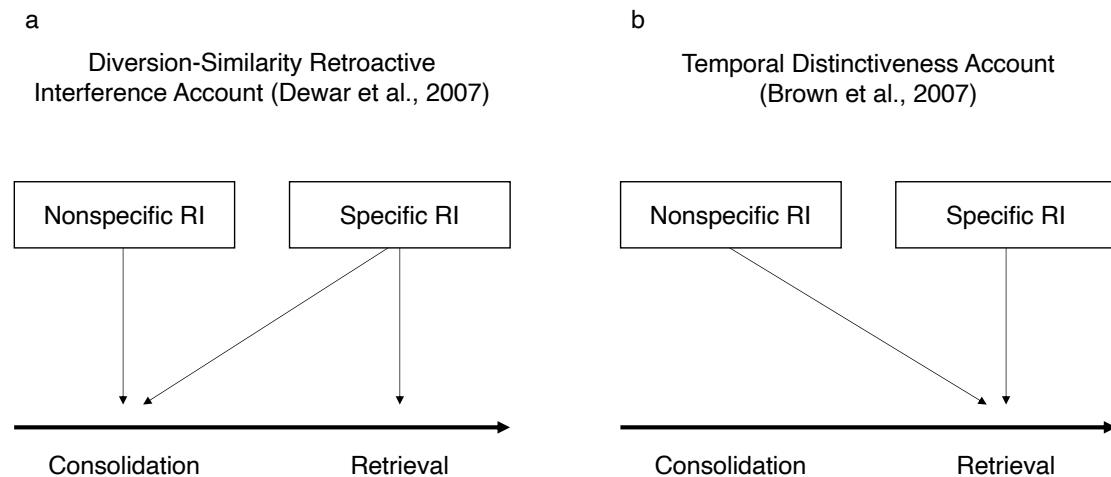
The general notion of consolidation being impaired by everyday nonspecific retroactive interference was developed further by Mednick et al. (2011) as an explanation of sleep-induced retrograde facilitation. According to their opportunistic consolidation hypothesis, the encoding of new information and the consolidation of previously created memory representations compete for limited hippocampal resources, that is, they are to some degree mutually exclusive (see also Wamsley, 2019). Whereas new encoding is prioritized over consolidation whenever the hippocampus is occupied by ongoing

processing demands, the minimization of such nonspecific retroactive interference is thought to free up hippocampal resources for consolidation. Thereby, opportunistic consolidation does not only explain sleep-induced retrograde facilitation in terms of uninterrupted consolidation due to the minimization of nonspecific retroactive interference, but predicts that consolidation should be enhanced by *any* activity or task that temporarily reduces encoding demands or abilities, such as, for example, acute alcohol consumption, relatively simple distractor tasks, or unoccupied rest.

Crucially, in the Mednick et al. (2011) framework, the opportunistic consolidation hypothesis is explicitly distinguished from two alternative accounts, that is, the unique-to-sleep consolidation hypothesis and the passive interference reduction hypothesis. Just as the opportunistic consolidation hypothesis, the unique-to-sleep consolidation hypothesis attributes sleep-induced retrograde facilitation to enhanced consolidation during sleep. However, these hypotheses differ in that the unique-to-sleep hypothesis assumes episodic memory consolidation to depend on specific neural mechanisms that are unique to slow wave sleep (SWS). The active systems consolidation account can be considered an elaborate specification of unique-to-sleep consolidation. According to this widely accepted interpretation of sleep-induced retrograde facilitation, neural replay (i.e., reactivation) of previously created memory traces in the hippocampus during SWS initiates the redistribution of the respective memory representations from the hippocampus to neocortical areas. Thereby, memory traces are not only strengthened, but also integrated into pre-existing long-term memory networks (see Rasch & Born, 2013). In contrast, the passive interference reduction hypothesis can be regarded as a representation of more traditional retroactive interference accounts. From this perspective, sleep-induced retrograde facilitation and other interference-based forgetting phenomena are driven by retrieval processes and depend on some rather specific (i.e., similarity-based) retroactive interference to occur during wakefulness.

Although the Mednick et al. (2011) terminology offers a useful framework for categorizing different accounts of interference-based forgetting, it suffers from an implicit “either-or” assumption, that is, opportunistic consolidation, unique-to-sleep consolidation, and passive interference reduction are treated as mutually exclusive. However, these hypotheses may instead be combined to predict and comprehensively explain various interference-based forgetting phenomena (see Berres, 2023).

Such a more integrative approach is implemented in a theoretical model proposed by Dewar et al. (2007). In line with a distinction of nonspecific and specific retroactive interference, these authors differentiate two different forms of retroactive interference,

Figure 1*Comparison of Interference-Based Accounts of Forgetting*

Note. Nonspecific retroactive interference (RI) is induced by any kind of interpolated cognitive processing and encoding demands, whereas specific RI is only induced by interpolated learning material similar to the original learning material. Depending on the theoretical account, these two types of RI are assumed to selectively affect the consolidation and/or retrieval of previously encoded memories.

that is, diversion and similarity retroactive interference. Whereas diversion retroactive interference is thought to be induced by any interpolated tasks and materials, and to inhibit consolidation, similarity retroactive interference is assumed to only be induced by similar interpolated tasks and materials, and to affect retrieval (see Figure 1a). Thus, the key prediction to be derived from this diversion-similarity account is that retroactive interference *always* affects consolidation, but only *sometimes* retrieval (i.e., whenever the similarity between original and interpolated materials is relatively high). Using the terminology introduced by Mednick et al. (2011), the diversion-similarity account integrates elements of opportunistic consolidation and passive interference reduction.

Despite the field's general focus on consolidation-based accounts of retroactive interference since Wixted (2004), theoretical accounts more in line with the traditional literature have been developed further as well. Most notably, Brown et al. (2007) proposed a temporal distinctiveness account that dispenses with any consolidation contributions to interference-based forgetting (see also Ecker et al., 2015). Specifically,

this account assumes episodic memory representations to be distributed within a multidimensional psychological space. This latent psychological space is defined by at least a temporal dimension along which memory representations are positioned according to their time of encoding. The basic assumption is that the retrieval probability of a specific memory representation is a direct function of its distinctiveness, that is, its temporal isolation with respect to other memory representations situated in its vicinity. In other words, any forgetting is attributed to retrieval problems due to low (temporal) distinctiveness from retroactive (or proactive) interference. Thus, the key prediction of the temporal distinctiveness theory is that retroactive interference *never* affects consolidation, but *always* retrieval (see Figure 1b). Thereby, it represents a revision of the passive interference reduction hypothesis as outlined by Mednick et al. (2011).

1.2 Empirical Findings on Interference-Based Forgetting

The opportunistic consolidation hypothesis has been motivated and justified by empirical work on the positive effects of various post-encoding activities and tasks that either reduce, delay, or minimize retroactive interference (see Mednick et al., 2011; Wixted, 2004). Three specific memory phenomena are typically treated as central evidence in favor of opportunistic consolidation: alcohol-induced retrograde facilitation, the temporal gradient of retroactive interference, and rest-induced retrograde facilitation.

The positive effect of post-encoding alcohol consumption on subsequent memory performance was first demonstrated by Parker et al. (1980, 1981). In their most rigorous experiment (Parker et al., 1981), the authors used a within-participants design that included one placebo and three alcohol conditions (low vs. medium vs. high dose). Participants studied a list of pictures before being administered their respective alcoholic or non-alcoholic beverage. After a 7-hr retention interval spent in the laboratory, memory was tested in a delayed recognition test. The authors observed a dose-dependent retrograde facilitation effect, that is, recognition performances in the medium and high dose conditions were significantly better than in the placebo condition. Later conceptual replications that employed different methodological approaches provided converging evidence for the robustness of the effect (e.g., Carlyle et al., 2017; Knowles & Duka, 2004; Mann et al., 1984; Weafer et al., 2016a, 2016b).

Alcohol-induced retrograde facilitation is typically seen as key evidence in favor of opportunistic consolidation because reduced encoding abilities during acute alcohol intoxication should protect hippocampal resources against nonspecific and specific

retroactive interference and thereby allow for consolidation processes to occur. Interestingly, Parker et al. (1981) already favored an active consolidation account to explain their result pattern but recognized the plausibility of a passive interference reduction account as well. Their main argument against a passive interference explanation was that participants did not engage in any intentional learning tasks during the retention interval. However, later studies found the retrograde facilitation effect to be more pronounced in more retrieval-dependent memory tests such as free recall as opposed to more retrieval-independent tests such as recognition (Mueller et al., 1983; Tyson & Schirmuly, 1994), a pattern that is more easily explained in terms of a passive interference reduction account. Against this backdrop, Wixted (2004) concluded that attempts to differentiate these accounts had “proven to be inconclusive” and that “a choice between them will probably depend on the identification of the specific physiological mechanism” (p. 255).

Instead of *reducing* retroactive interference by means of alcohol consumption, interpolated learning may also be *delayed* to enhance subsequent memory performance. More specifically, early work by Müller and Pilzecker (1900) already suggested that retroactive interference follows a temporal gradient. In one of their experiments, a single participant repeatedly studied and recalled lists of syllable pairs. Crucially, in some trials, the time interval between the original and an interpolated list lasted 6 minutes, in others only 17 seconds. The authors observed higher cued recall rates in those trials with longer time intervals. Although later research on the temporal gradient of retroactive interference yielded rather mixed results (Wickelgren, 1977), Wixted (2004) convincingly argued that most replication failures could likely be attributed to methodological problems. Indeed, more recent experiments consistently found a temporal gradient effect (Ecker et al., 2015; Mercer, 2015).

As for alcohol-induced retrograde facilitation, the identification of the precise mechanisms underlying the temporal gradient of retroactive interference continue to be contested. Indeed, Wixted (2004) based his initial proposal of consolidation-based retroactive interference mainly on temporal gradient effects. One crucial assumption of the opportunistic consolidation hypothesis is that early consolidation processes render the respective memory trace less susceptible to retroactive interference. Thus, the longer the time interval between original and interpolated learning, the higher the probability that the memory trace will already be stable enough to endure the damaging influence of retroactive interference.

Having said that, the temporal distinctiveness account provides a more parsimo-

nious explanation of the temporal gradient of retroactive interference without the need to invoke some consolidation mechanism. Instead, it attributes the effect to higher versus lower temporal isolation of the original learning material with respect to the interpolated material. In line with this reasoning, Ecker et al. (2015) found that for their computational implementation of the temporal distinctiveness account, model fit indices favored model versions without a consolidation mechanism. However, more direct evidence for and against both theoretical accounts has not been reported. To illustrate, in line with Wixted's (2004) assessment of alcohol-induced retrograde facilitation, Mercer (2015) had to concede that his experiment "cannot disentangle consolidation and distinctiveness-based accounts since both models predict that postponing RI [retroactive interference] will reduce forgetting" and that future attempts would require "an ingenious design to fully extricate the predictions of these two accounts" (p. 134).

A third line of research has used short periods of post-encoding waking rest to effectively *minimize* both specific and nonspecific retroactive interference instead of merely reducing or delaying it. Indeed, a growing body of evidence suggests that a few minutes of eyes-closed, unoccupied rest can facilitate subsequent declarative and procedural memory performances (Wamsley, 2019; but see Martini & Sachse, 2020). Notably, such effects have even been observed when using rather dissimilar distractor tasks that seem unlikely to induce any specific retroactive interference. For example, in a study reported by Dewar et al. (2012), participants listened to a short story before either wakefully resting or engaging in a purely visual spot-the-difference game for 10 minutes. Importantly, the short story and spot-the-difference game did not share any semantic overlap, that is, specific retroactive interference was most likely minimized in both conditions. Nevertheless, participants' immediate and delayed story recall was significantly better in the waking rest than in the distractor condition.

At first glance, such demonstrations of very unspecific retroactive interference could be interpreted as evidence against a passive interference reduction account. In line with this, Wamsley (2019) interpreted the available evidence in favor of an opportunistic consolidation account while acknowledging that the reported studies "suggest (but do not prove) that the effect of rest on memory is not due to a simple reduction in sensory interference" (p. 172). However, the temporal distinctiveness account provides a straightforward explanation for such observations based on the idea that even rather dissimilar interpolated memory representations may decrease the temporal isolation of the to-be-remembered information. In other words, from a distinctiveness perspective,

similarity is not assumed to be a necessary prerequisite for retroactive interference to occur, but rather one of many dimensions that all contribute to higher versus lower overall distinctiveness and retrievability. Thus, although the waking rest effect is typically treated as key behavioral evidence in favor of opportunistic consolidation, it may also be explained through temporal distinctiveness.

Overall, there is much behavioral evidence that implies memory benefits resulting from reduced, delayed, or minimized retroactive interference during wakefulness. However, interpretations of such observations as evidence in favor of opportunistic consolidation are premature as long as no direct evidence exists that would allow for a clear-cut differentiation of opportunistic consolidation from passive interference reduction such as suggested by the temporal distinctiveness account. The application of storage-retrieval MPT models offers an opportunity for such a differentiation.

2 Storage-Retrieval Analyses

Generally speaking, the successful recollection of some previously encountered piece of information necessarily presupposes that this information is not only available in memory, but also retrievable. Thus, a correct response in a recall test implies that both storage and retrieval of the respective information must have been successful. Conversely, however, an incorrect response can be attributed to either unsuccessful storage (i.e., the information was not encoded in the first place or was not maintained in memory across the retention interval) or unsuccessful retrieval (i.e., the information was stored successfully but could not be retrieved). For memory tests that do not impede successful guessing, the interplay of latent mechanisms underlying correct and incorrect responses becomes even more complex: For example, in a typical old-new recognition test, a correct response may be the result of successful storage and retrieval, or correct guessing in a state of uncertainty. Conversely, an incorrect response may imply either unsuccessful storage and incorrect guessing, or successful storage, unsuccessful retrieval, and incorrect guessing. Therefore, surface measures of memory such as the number of correct responses in recall tests or hit and false-alarm rates in recognition tests entail a high degree of uncertainty with respect to the underlying cognitive mechanisms.

Crucially, the opportunistic consolidation and the temporal distinctiveness account of interference-based forgetting clearly differ with respect to the mechanism supposedly targeted by retroactive interference. On the one hand, the opportunistic consolidation account assumes that memory traces are stabilized against disruptive influences during periods of reduced retroactive interference. Thus, retroactive interference should impede the enduring *storage* of the to-be-remembered information in memory. The assumption that retrieval should not be affected directly by nonspecific retroactive interference is expressed in the Dewar et al. (2007) diversion-similarity model. On the other hand, the temporal distinctiveness account incorporates and emphasizes the traditional view of retroactive interference as a *retrieval* phenomenon without any contribution of consolidation (Brown et al., 2007; Tulving & Psotka, 1971). Thus, a severe and fair test of the opportunistic consolidation and the temporal distinctiveness account requires an effective approach to precisely disentangle storage (i.e., consolidation) and retrieval contributions to interference-based forgetting.

So far, attempts to differentiate opportunistic consolidation from temporal distinctiveness have been scarce. One of the few exceptions is the observation that

alcohol-induced retrograde facilitation seems to be more pronounced in more retrieval-dependent memory tests such as free recall compared to more retrieval-independent tests such as recognition (Mueller et al., 1983; Tyson & Schirmuly, 1994). For example, Mueller et al. (1983) found a significant alcohol versus placebo effect on delayed free recall performance for word lists but no significant effect on delayed recognition performance. Such a result pattern suggests that the effect was retrieval-driven, without a contribution of storage processes.

Comparisons of effects on differently retrieval-dependent memory tests between experimental conditions can indeed provide some tentative insights into the underlying mechanisms (see also Drachman & Leavitt, 1972; Hogan & Kintsch, 1971). More specifically, if an experimental manipulation has an effect on the more retrieval-dependent memory test (e.g., free recall) but not on the more retrieval-independent test (e.g., recognition), retrieval processes can be assumed to play a major role. In contrast, if the result pattern entails an effect on both memory tests, it remains unclear whether the effect was driven by storage, or both storage and retrieval. Thus, despite their appealing simplicity, such analyses of performance profiles seem insufficient to arrive at definite conclusions for all possible effect combinations (see Küpper-Tetzel & Erdfelder, 2012).

2.1 Multinomial Processing Tree Modeling

More fine-grained conclusions can be derived by retaining the basic idea of comparing effects on differently retrieval-dependent memory tests but adopting a more sophisticated MPT analysis approach (see Erdfelder et al., 2009 for a review of applications; see Schmidt et al., 2023 for a tutorial). MPT models form a class of stochastic models that are tailored to the analysis of categorical data from specific experimental paradigms. They allow researchers to disentangle and measure the latent contributions of intertwined cognitive processes that underlie behavioral responses (Batchelder & Riefer, 1999; Riefer & Batchelder, 1988). In the past, MPT models have been successfully applied to a wide range of paradigms and substantive research questions in many subdisciplines of psychology such as attention and perception, learning and memory, judgment and decision making, and social cognition (Schmidt et al., 2023).

Storage-retrieval MPT models allow researchers to precisely disentangle latent storage, retrieval, and (if applicable) guessing contributions to directly observable performances on surface memory measures. Such models have been developed for various experimental paradigms including free-then-cued-recall (Küpper-Tetzel &

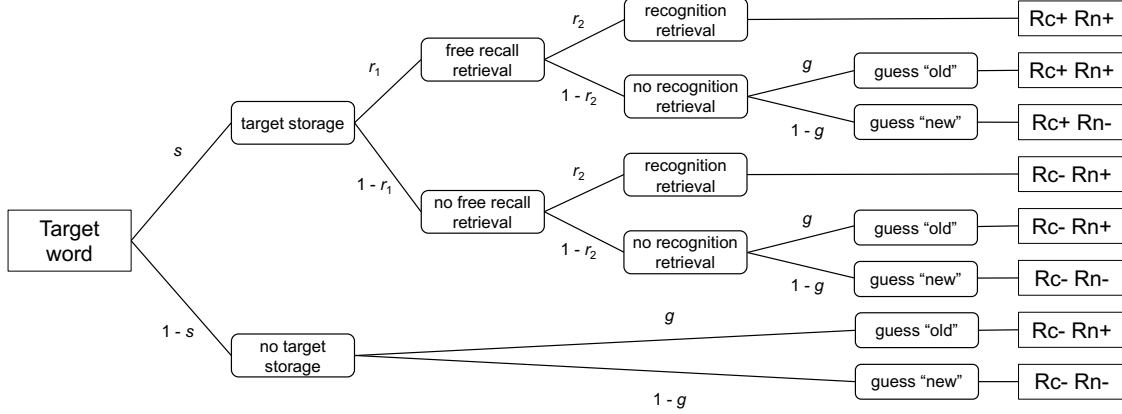
Erdfelder, 2012; Rouder & Batchelder, 1998) and recognition-then-cued-recall (Riefer & Batchelder, 1995) testing procedures for word pairs as learning material. Crucially, such previously validated and applied models may be adapted, for example, to allow for a more fine-grained differentiation of encoding and maintenance instead of a single storage parameter (see Manuscript I, see also Küpper-Tetzl & Erdfelder, 2012) or to accommodate other types of learning materials or testing procedures (see Manuscripts II and III).

To illustrate, consider the free-recall-then-recognition storage-retrieval MPT model used in Experiments 2 and 3 of Manuscript II. This model is an adapted version of the recognition-then-cued-recall model proposed by Riefer and Batchelder (1995). The combination of responses on a free recall and a subsequent recognition test for single words as learning material yields a 2 x 2 matrix of possible response combinations, that is, in both tests a response may be correct or incorrect. To infer the respective contributions of storage and retrieval processes, the probabilities of the four response categories are reparameterized by means of a set of latent model parameters. These include one storage parameter s , two retrieval parameters r_1 (for free recall) and r_2 (for recognition), and a guessing parameter g to represent the probability of guessing “old” in the recognition test in case of storage or retrieval failure. Category probabilities are expressed as functions of these latent model parameters such that sequences of processing steps or parameters (so-called “branches”) may result in identical or different observations (i.e., each branch terminates in exactly one category, but multiple branches may terminate in the same category). Assuming that parameters within branches represent probabilities that are conditional on previous parameters of that branch, model equations are obtained by multiplying parameters within branches and adding the products of parameter sequences terminating in the same category. For example, reproducing a target word during free recall and correctly classifying it as being “old” during recognition may be assumed to be the result of either successful storage, successful free recall retrieval, and successful recognition retrieval (i.e., $s \cdot r_1 \cdot r_2$), or successful storage, successful free recall retrieval, unsuccessful recognition retrieval, and correct guessing (i.e., $s \cdot r_1 \cdot (1 - r_2) \cdot g$). Thus, the probability of this response category is equal to the sum of these two branches, that is, $s \cdot r_1 \cdot r_2 + s \cdot r_1 \cdot (1 - r_2) \cdot g$. The target tree structure of this model is illustrated in Figure 2.

Model fit indices and parameter estimates can be obtained from a range of different estimation approaches (Singmann et al., 2024). Traditionally, MPT models have been fitted by means of maximum likelihood (ML) estimation based on category

Figure 2

Target Tree of the Storage-Retrieval MPT Model from Manuscript II



Note. Multinomial processing tree (MPT) parameter s = probability of successfully storing a target word in memory, r_1 = probability of successfully retrieving a stored target word during free recall, r_2 = probability of successfully retrieving a stored target word during recognition, g = probability of guessing "old" during recognition given no recognition retrieval or distractor detection. Rc+ = successful target recall, Rc- = unsuccessful target recall, Rn+ = successful target recognition, Rn- = unsuccessful target recognition. For the distractor tree structure, see Manuscript II.

frequencies aggregated across participants (Moshagen, 2010; Singmann & Kellen, 2013). Thereby, this *complete pooling* approach rests on the assumption of identically and independently distributed (i.i.d.) observations. In other words, parameters are expected not to vary between participants, an assumption that seems rather questionable and may result in incorrect statistical inferences (Schmidt et al., 2023; Smith & Batchelder, 2008). As an alternative, *partial pooling* estimation approaches have been developed that account for inter-individual differences. For example, the Bayesian hierarchical latent-trait approach (Klauer, 2010) assumes individual parameters to follow a multivariate normal distribution at the group level. Thereby, correlations between model parameters may also be estimated. In all three manuscripts included in this thesis, both estimation approaches were used to ensure the robustness of our model-based conclusions. Overall, our results largely proved not to be sensitive to the differential distributional assumptions of both approaches, a finding that is in line with a recent meta-analysis (Singmann et al., 2024; for an analysis of aggregation invariance of MPT models, see Erdfelder et al., in press).

2.2 Sleep-Induced Retrograde Facilitation

By adopting a storage-retrieval MPT approach, behavioral data from relevant memory experiments can be used to effectively differentiate consolidation from distinctiveness explanations of interference-based forgetting. To illustrate, in two experiments, Erdfelder et al. (2022) employed an MPT modeling approach to disentangle storage and retrieval contributions to memory performance following a 12-hr retention interval spent either awake or mostly asleep. In both experiments, participants were presented with a list of 40 weakly associated word pairs during the learning phase. This was followed by an immediate cued recall test in which participants were presented with the first (i.e., cue) word of each pair and asked to recall the corresponding second (i.e., target) word. After a retention interval, memory was tested by means of a final free-then-cued-recall procedure. In the free recall test, participants were given 8 minutes to freely recall as many of the previously studied word pairs as possible in any order. They were instructed to write down single words as well if they could not remember the entire word pair. The subsequent cued recall test was identical to the immediate cued recall test.

Such a procedure involving two cued recall tests (immediate and final) and a free recall test allows for a model-based analysis of the resulting recall data by means of the encoding-maintenance-retrieval MPT model introduced by Küpper-Tetzl and Erdfelder (2012). Thereby, encoding, maintenance, and retrieval contributions to participants' recall performances can be precisely disentangled. With respect to sleep-induced retrograde facilitation, maintenance parameter m should be reflective of unique-to-sleep or opportunistic consolidation contributions, whereas free recall retrieval parameter r_f should be reflective of temporal distinctiveness or other passive interference reduction contributions.

In Experiment 1, we used a 2 x 2 between-participants design with the factors "Study Time" (morning vs. evening) and "Retention Interval" (6 minutes vs. 12 hours). The sample size was $N = 40$. Participants in the 12-hr retention interval condition spent the time between sessions outside the laboratory, and, depending on the study time, either pursuing everyday activities (wake condition) or spending the night asleep (sleep condition). The model-based results revealed significantly higher maintenance probabilities m and a descriptive trend towards higher free recall retrieval probabilities r_f in the sleep compared to the wake condition. Moreover, encoding probabilities e remained unaffected by both experimental factors, suggesting that sleep-induced retrograde facilitation cannot be attributed to circadian rhythms of encoding ability.

We followed up on our observations from Experiment 1 in a second experiment that involved a larger sample size ($N = 60$) and an experimental manipulation intended to provide evidence for the differentiation of parameters m and r_f . To this end, we again used a 2 x 2 between-participants design in Experiment 2, this time with the factors “Study Time” (morning vs. evening) and “Retrieval Cues” (absent vs. present). Our expectation was that the presentation of category labels as retrieval cues during the free recall test should have a selective influence on free recall retrieval probabilities r_f but not on maintenance probabilities m . In line with our hypotheses, we replicated the positive effects of sleep on both parameters m and r_f . Importantly, the presentation of retrieval cues had no effect on parameter m but only on parameter r_f . Thus, the dual sleep benefit for both maintenance and retrieval cannot be explained in terms of an imprecise differentiation of these processes in the MPT model.

Across both experiments, we found convincing evidence for both passive interference reduction and active consolidation explanations of sleep-induced retrograde facilitation. Crucially, however, the model-based results are necessarily inconclusive with respect to the differentiation of unique-to-sleep and opportunistic consolidation since both accounts predict a positive effect of sleep on storage or maintenance probabilities. Instead, somewhat counterintuitively, memory experiments without any involvement of sleep are needed for this purpose.

In fact, some applications of a storage-retrieval MPT model to the investigation of specific retroactive interference during wakefulness have already been reported in the literature (Bäuml, 1991a, 1991b, 1991c; Riefer & Batchelder, 1988). In all of these studies, the Tulving and Psotka (1971) paradigm or variations thereof were adapted to accommodate the application of a specific storage-retrieval model, that is, the pair-clustering model by Batchelder and Riefer (1980). The general conclusion from this research was that, in line with the initial proposal by Tulving and Psotka (1971), the number of interpolated word lists primarily affects retrieval of the original list. Under some rather specific conditions, however, storage may be impaired as well (see Bäuml, 1991a, 1991b). Unfortunately, the results from this line of research are hardly applicable to the differentiation of the opportunistic consolidation and the temporal distinctiveness account.

In the research reported in the manuscripts included in this thesis, we replicate relevant interference-based forgetting effects that occur during wakefulness, and re-evaluate them in terms of underlying storage and retrieval processes by means of appropriate storage-retrieval MPT models. Thus, the main goal of this thesis is to

disentangle opportunistic consolidation and temporal distinctiveness contributions to interference-based forgetting during wakefulness. At the same time, the results reported in the three manuscripts may also be used for a unique-to-sleep versus opportunistic consolidation interpretation of the storage effect we found to underlie sleep-induced retrograde facilitation. Thereby, this thesis contributes to a comprehensive understanding of interference-based forgetting during both wakefulness and sleep.

3 Re-Evaluation of Behavioral Evidence

In each of the three manuscripts included in this thesis, we intended to replicate one of the empirical findings typically treated as key evidence in favor of opportunistic consolidation, that is, alcohol-induced retrograde facilitation (Manuscript I), the temporal gradient of retroactive interference (Manuscript II), and rest-induced retrograde facilitation (Manuscript III), and re-evaluate the resulting data by means of a suitable storage-retrieval MPT model. In the following, I outline the most important aspects of each manuscript with respect to the methodological approach and the main results.

3.1 Alcohol-Induced Retrograde Facilitation

Quevedo Pütter, J., & Erdfelder, E. (2022). Alcohol-induced retrograde facilitation? Mixed evidence in a preregistered replication and encoding-maintenance-retrieval analysis. *Experimental Psychology*, *69*(6), 335-350. <https://doi.org/10.1027/1618-3169/a000569>

In Manuscript I, we report an experiment that was designed to replicate as closely as possible the alcohol-induced retrograde facilitation effect found by Parker et al. (1981) while at the same time allowing for an application of the Küpper-Tetzel and Erdfelder (2012) free-then-cued-recall storage-retrieval MPT model used by Erdfelder et al. (2022). Most importantly, just as Parker et al. (1981), we used a 7-hr retention interval, whereas the authors of other replication studies had either employed considerably shorter retention intervals (Knowles & Duka, 2004; Mann et al., 1984; Parker et al., 1980, study 1; Tyson & Schirmuly, 1994) or had extended the retention interval up to 48 hours by having participants spend the time interval between post-encoding alcohol versus placebo administration and final memory testing outside the laboratory (Bruce & Pihl, 1997; Carlyle et al., 2017; Lamberty et al., 1990; Mueller et al., 1983; Parker et al., 1980, study 2; Weafer et al., 2016a, 2016b). Such procedural adaptations entail the risk of unintended confounding influences of ongoing alcohol intoxication during memory testing or alcohol-induced changes to the sleep architecture. Thus, our study was the first to put the retrograde facilitation effect found by Parker et al. (1981) to a methodologically rigorous test.

For this purpose, a total of $N = 93$ participants took part in an extensive laboratory experiment that included an initial cued recall and final free-then-cued-recall testing procedure within a single experimental session, using the same learning material that

we already used in both experiments on sleep-induced retrograde facilitation (Erdfelder et al., 2022) described in the Storage-Retrieval Analyses section. Participants were randomly assigned to an alcohol or a placebo condition and all participants received their respective alcoholic or non-alcoholic beverage immediately following the initial cued recall. The alcoholic beverage was expected to result in peak blood alcohol concentrations of around 0.60‰. Breath alcohol concentrations were measured 30, 60, and 90 minutes after the end of the alcohol versus placebo administration, and again immediately before the final free recall test. Participants were encouraged to register in groups and spent the 7-hr retention interval in a seminar room where they watched a standardized series of movies and were free to interact among each other. During this time, participants were supervised to ensure that they did not fall asleep or explicitly discuss any of the previously encoded learning material. The data was collected and analyzed by means of a sequential testing procedure (i.e., the sequential probability ratio t test, Schnuerch & Erdfelder, 2020) to maximize efficiency.

We found no significant differences between conditions in either surface memory measure, that is, we did not replicate the alcohol versus placebo effect on either cued recall retention or free recall performance. However, our model-based results provide evidence for a retrieval benefit in the alcohol condition. More specifically, free recall retrieval parameter r_f was estimated to be larger in the alcohol compared to the placebo condition. This clear descriptive pattern was reliable for both estimation approaches. In contrast, there was only a very small descriptive difference between conditions for maintenance parameter m and this pattern was only reliable for the aggregated but not the individual data. As expected, encoding parameter e did not differ reliably between conditions either and cued recall retrieval parameter r_c was estimated to be very close to 1 in both conditions. Estimates for these key parameters are provided in Table 1.

This pattern of surface and model-based results suggests that alcohol-induced retrograde facilitation is (a) less robust than suggested by previous studies that did not use the original 7-hr retention interval and (b) driven by retrieval rather than storage or maintenance processes. Thereby, Manuscript I provides direct evidence in favor of a passive retroactive interference reduction and against an opportunistic consolidation account of alcohol-induced retrograde facilitation. Importantly, in line with the temporal distinctiveness account and the results on sleep-induced retrograde facilitation by Erdfelder et al. (2022), our results from Manuscript I suggest that nonspecific retroactive interference can be sufficient to impair subsequent retrieval.

Table 1*Results of the Encoding-Maintenance-Retrieval MPT Analysis in Manuscript I*

Parameter	Alcohol	Placebo
	Aggregated data ^a	
e	.58 [.55, .61]	.60 [.57, .62]
m	.92 [.90, .95]	.89 [.86, .91]
r_f	.50 [.46, .54]	.44 [.41, .47]
r_c	.98 [.97, .98]	.98 [.97, .98]
	Individual data ^b	
e	.59 [.53, .64]	.60 [.55, .65]
m	.93 [.90, .96]	.90 [.86, .93]
r_f	.51 [.45, .56]	.43 [.39, .48]
r_c	.98 [.97, .99]	.98 [.97, .99]

Note. Multinomial processing tree (MPT) parameter e = probability of associative encoding of a word pair, m = probability of associative maintenance of a word pair across the retention interval, r_f = probability of retrieving a word pair as an association during free recall, r_c = probability of retrieving a word pair as an association during cued recall. For the remaining parameter estimates, please refer to Manuscript I.

^a The model was fitted to the aggregated category frequencies using maximum likelihood (ML) estimation in the multiTree software (Moshagen, 2010). Parameter estimates are presented alongside the corresponding 95% confidence intervals.

^b The model was fitted to the individual category frequencies using Bayesian hierarchical estimation in the R package TreeBUGS (Heck et al., 2018). Posterior means are presented alongside the corresponding 95% Bayesian credibility intervals.

Some limitations of this study should be acknowledged. First, the application of the Küpper-Tetzel and Erdfelder (2012) encoding-maintenance-retrieval MPT model required word pairs instead of single items as learning material. Thus, in contrast to the original Parker et al. (1981) experiment, our replication study investigated associative instead of item memory. Although the passive interference reduction and the opportunistic consolidation account of interference-based forgetting do not explicitly

predict that the susceptibility to retroactive interference should differ according to the associative or non-associative nature of the learning material, such a prediction can be derived from the memory-system dependent forgetting hypothesis (Hardt et al., 2013; Kuhlmann et al., 2021). According to this hypothesis, hippocampally represented associative memories should be less susceptible to retroactive interference than extra-hippocampally represented item memories. Thus, had we used single items instead of word pairs as learning material, the retrieval benefit we observed for the alcohol condition might have been even more pronounced, possibly resulting in a significant surface-level free recall effect.

Second, we observed lower-than-expected mean peak breath alcohol concentrations in the alcohol condition, that is, $M = 0.43\text{‰}$. Thus, it could be argued that encoding abilities of participants in the alcohol condition were not impaired sufficiently to result in a clear retroactive interference difference between conditions. Indeed, lower peak alcohol concentrations also imply shorter latencies until participants were completely sober again. Assuming an alcohol elimination rate of 0.15‰ per hour (Thierauf et al., 2013), most participants in the alcohol condition must have been sober after about 3 hours. During the remaining 4 hours of the retention interval, equal degrees of retroactive interference can therefore be expected in both conditions. Although the mean peak breath alcohol concentration in our study was still higher than in the medium-dose condition by Parker et al. (1981), such concerns appear quite reasonable and should be considered in future studies.

Third, the very nature of the experimental manipulation implies a rather low experimental control in both the alcohol and the placebo condition during the 7-hr retention interval. Thereby, it remains unclear whether the alcohol versus placebo administration caused behavioral and cognitive differences between conditions beyond a mere reduction in encoding abilities in the alcohol condition. For example, acute alcohol consumption has also been shown to increase mind-wandering (Sayette et al., 2009), a spontaneous cognitive activity that has been argued to be sufficient to interfere with consolidation (Humiston et al., 2019). Thus, alcohol-induced retrograde facilitation might be the result of a potentially large number of both positive and negative memory-relevant factors affected by acute alcohol consumption. Importantly, such confounding variables should be even more relevant if participants leave the supervised laboratory between experimental sessions. Nevertheless, a more controlled procedural approach is needed to more precisely identify the unique contributions of specific and nonspecific retroactive interference on subsequent memory performance.

3.2 Temporal Gradient of Retroactive Interference

Quevedo Pütter, J., Dahler, S., & Erdfelder, E. (2024). *Opportunistic consolidation or temporal distinctiveness? Retrieval, not storage, drives the temporal gradient of retroactive interference in episodic memory*. Manuscript submitted for publication.

Instead of *reducing* nonspecific and specific retroactive interference through alcohol consumption, retroactive interference may also be *delayed* to benefit subsequent memory performance. In Manuscript II, we report a series of three experiments that were designed to scrutinize the replicability of the temporal gradient of retroactive interference (TGRI) and to disentangle its underlying mechanisms.

Our methodological approach in Manuscript II alleviates many of the limitations associated with the experiment reported in Manuscript I. First, instead of adapting the to-be-replicated original study to accommodate the application of a certain MPT model, we adapted an existing storage-retrieval model proposed by Riefer and Batchelder (1995) to allow for a very close replication of a study reported by Ecker et al. (2015). Thereby, we could adopt the pool of single words used by these authors instead of word pairs to increase the chances of observing significant effects on free recall performances in line with the memory-system dependent forgetting hypothesis (Kuhlmann et al., 2021). Moreover, the use of a simple distractor task during the retention interval and intentional rather than implicit learning instructions for the interpolated learning task ensured a high experimental control in all conditions. Indeed, by manipulating the timing rather than the amount of interpolated learning, unintended influences of confounding variables should become rather unlikely.

In Experiment 1, we conducted a very close replication of the first of two studies reported by Ecker et al. (2015). To this end, $N = 80$ participants took part in an online experiment. Across eight experimental trials, participants studied an original list of 10 words (L1), an interpolated list of 10 additional words (L2), and were tested on a free recall test for the L1 words (T), followed by another free recall for the L2 words in 50% of all trials. We used a within-participants manipulation of the timing of the interpolated L2 learning phase: In LS trials, the L1-L2 intervals lasted 240 seconds and the L2-T interval lasted 60 seconds, whereas in SL trials, these durations were reversed. During the L1-L2 and L2-T intervals, participants worked on a simple color-detection distractor task. Blue target and grey distractor squares were presented sequentially and randomly intermixed in the middle of the computer screen. Participants were asked to press a key as quickly as possible whenever a target

square was presented but not to react to distractor squares. This task was designed to replicate as closely as possible the tone-detection task used by Ecker et al. (2015). We included the original data in our analysis to perform a 2 x 2 mixed ANOVA with the factors “Study” (original vs. replication) and “L2 Timing” (LS vs. SL), and the number of correct responses in the free recall as our dependent variable. Thereby, a successful replication would be indicated by a non-significant interaction effect (see Anderson & Maxwell, 2016).

Descriptively, participants in our replication study performed better in the LS than in the SL condition. In line with this descriptive difference, the within-between interaction effect turned out to be non-significant. Thus, the direction and size of the LS versus SL effect in our study was consistent with the effect observed by the original authors.

Based on the replication success in Experiment 1, we aimed at disentangling storage and retrieval contributions to the TGRI in Experiment 2. Therefore, it was necessary to extend the original procedure to include an additional memory test that was less retrieval-dependent than the free recall test, and to specify a storage-retrieval MPT model specifically tailored to such a paradigm. Since no such model for single items was readily available in the literature, we chose to adapt a model originally proposed by Riefer and Batchelder (1995) and developed further by Nadarevic (2017). The resulting model is tailored to a free-recall-then-recognition paradigm and allows to disentangle storage (parameter s), free recall retrieval (parameter r_1), recognition retrieval (parameter r_2), and guessing contributions (parameter g) to item memory. Since the inclusion of an old-new recognition test immediately *after* the respective L1 free recall test is unlikely to cause any bias in the current or subsequent trials, the results from Experiment 2 also served as an additional replication of the TGRI in free recall. Experiment 2 entailed only four instead of eight trials to keep the total study duration within a reasonable range. L1 recognition and L2 free recall tests were included in all trials. The final sample size was $N = 177$.

We found significantly better L1 free recall performances in the LS than the SL condition. In contrast, L1 recognition performances did not differ significantly between conditions. In line with this pattern, our model-based analyses revealed a clear free recall retrieval benefit in the LS condition, that is, parameter r_1 was estimated to be higher in the LS than in the SL condition. This pattern was reliable for both estimation approaches. In contrast, there was no evidence for a difference in storage probabilities s between conditions.

These results from Experiment 2 suggest that the temporal distinctiveness account provides a better explanation of the TGRI in free recall than the opportunistic consolidation account. This interim conclusion converges with the observation of better retrieval following post-encoding alcohol consumption in Manuscript I. That being said, proponents of the opportunistic consolidation account might argue that the uncontrolled online setting of Experiment 2 in Manuscript II did not allow for consolidation to occur during the L1-L2 and L2-T intervals since external distractions could not be prevented. Thus, although the results imply a role for temporal distinctiveness in the TGRI, the possibility of an additional contribution of opportunistic consolidation under more controlled conditions cannot be excluded. Therefore, the first aim of Experiment 3 was to replicate the results from Experiment 2 in a laboratory setting.

The interpretation of the results from Experiment 2 is based on the assumption that the stabilization of labile memory traces through opportunistic consolidation within the first minutes after encoding should primarily benefit memory storage. However, according to the opportunistic consolidation account, the integration into pre-existing memory networks should be initiated shortly after encoding as well. Such a qualitative transformation of memory traces might be expected to not only benefit memory storage but also retrieval thanks to the creation of new retrieval cues. From such a perspective, a retrieval benefit in the LS compared to the SL condition might be interpreted as evidence in favor of opportunistic consolidation. Importantly, such an interpretation would only be justified in case of a simultaneous storage *and* retrieval effect.

Against this backdrop, our second aim in Experiment 3 was to differentiate temporal distinctiveness and opportunistic consolidation explanations for a potential free recall retrieval benefit in the LS condition in case of a simultaneous storage effect. To this end, we used a 2 x 2 design with the factor “L2 Timing” (LS vs. SL) and the additional factor “L1-L2 Similarity” (high vs. low). We generated a pool of geometric figures as L2 learning material for the low L1-L2 similarity condition. Importantly, the temporal distinctiveness account acknowledges the existence of additional dimensions apart from the temporal one, such as a semantic dimension (Brown et al., 2007). Thus, in line with traditional retroactive interference accounts (see McGeoch & McDonald, 1931), such a *generalized* distinctiveness account would predict only a very small or even nonexistent LS versus SL effect on free recall retrieval in case of minimal L1-L2 similarity. Conversely, the opportunistic consolidation account predicts that any encoding demands should inhibit consolidation regardless of L1-L2 similarity,

Table 2*Results of the Storage-Retrieval MPT Analysis of Experiment 3 in Manuscript II*

Parameter	High similarity		Low similarity	
	LS	SL	LS	SL
Aggregated data ^a				
s	.84 [.82, .85]	.83 [.82, .85]	.85 [.83, .86]	.84 [.82, .85]
r_1	.70 [.68, .72]	.64 [.62, .66]	.70 [.68, .72]	.66 [.64, .68]
r_2	.99 [.98, .99]	.99 [.98, .99]	.99 [.99, .99]	.99 [.99, .99]
Individual data ^b				
s	.88 [.85, .90]	.86 [.84, .89]	.88 [.85, .90]	.86 [.84, .89]
r_1	.71 [.66, .75]	.65 [.61, .70]	.72 [.67, .76]	.67 [.63, .71]
r_2	.99 [.98, .99]	.99 [.98, .99]	.99 [.99, .99]	.99 [.99, .99]

Note. Multinomial processing tree (MPT) parameter s = probability of storage of an L1 word, r_1 = probability of retrieving an L1 word during free recall, r_2 = probability of retrieving an L1 word during recognition. For the estimates of guessing parameter g , please refer to Manuscript II.

^a The model was fitted to the aggregated category frequencies using maximum likelihood (ML) estimation in the R package MPTinR (Singmann & Kellen, 2013). Parameter estimates are presented alongside the corresponding 95% confidence intervals.

^b The model was fitted to the individual category frequencies using Bayesian hierarchical estimation in the R package TreeBUGS (Heck et al., 2018). Posterior means are presented alongside the corresponding 95% Bayesian credibility intervals.

so this factor should have no effect on the size of an LS versus SL effect on free recall retrieval. Put differently, the generalized distinctiveness account predicts an interaction effect of L2 timing and L1-L2 similarity on free recall retrieval, whereas the opportunistic consolidation account does not. To reiterate, this theoretical reasoning would become necessary only if a simultaneous storage and retrieval effect of the LS versus SL manipulation would be observed.

Experiment 3 took place in a controlled laboratory setting and included eight trials, that is, two trials per condition. For a closer replication of the original study by Ecker et al. (2015), we switched from the color-detection task used in Experiments 1 and 2 to the original tone-detection task. To increase the potential LS versus SL effect size, the interpolated L2 learning occurred even earlier in the SL condition than in the previous two experiments, that is, after 20 instead of 60 seconds. The final sample size was $N = 140$.

With respect to our surface memory measures, we found a significant main effect of L2 timing on L1 free recall performance. All other main and interaction effects on L1 free recall and recognition performances were non-significant. In line with this pattern, our model-based analyses revealed a reliable L2 timing main effect on free recall retrieval parameter r_1 , but no effect on storage parameter s . This result pattern is in line with our observation from Experiment 2, that is, we again found a TGRI in free recall that was purely retrieval-driven. Surprisingly, however, L1-L2 similarity had no effect on any of our surface and model-based memory measures. Thus, our results are in line with a purely temporal distinctiveness account but neither with the opportunistic consolidation nor a generalized distinctiveness account. As expected, recognition retrieval probabilities r_2 were estimated to be very close to 1. All storage and retrieval MPT parameter estimates are provided in Table 2.

Across all three experiments, our results from Manuscript II suggest (a) that the TGRI in free recall is a robust effect and (b) that it constitutes a retrieval rather than a storage phenomenon, regardless of the similarity or specificity of the interpolated material. Together with the results from Manuscript I, these observations suggest that both specific and nonspecific retroactive interference may result in retrieval impairments. Thereby, these results are at odds with the opportunistic consolidation account of interference-based forgetting.

From an opportunistic consolidation perspective, two characteristics of all four experiments reported in Manuscripts I and II might have contributed to our consistent null-findings with respect to storage processes. First, we used comparatively short retention intervals of 7 hours (Manuscript I) and 5 minutes (Manuscript II) to conduct all experiments within one single session. Although opportunistic consolidation has been argued to set in immediately after encoding (Wamsley, 2019) and should therefore influence memory performances even on rather short time scales, researchers have highlighted the supposed long-term effects of opportunistic consolidation on longer time scales of several days (e.g., Dewar et al., 2012; Martini et al., 2020).

Second, although retroactive interference was either reduced (Manuscript I) or delayed (Manuscript II), it has been argued that opportunistic consolidation requires conditions of *minimal* retroactive interference (Dewar et al., 2007; Wixted, 2004). From the perspective of such a strict interpretation of the opportunistic consolidation account, acute alcohol intoxication or the attentional focus on a simple distractor task is insufficient to spare hippocampal resources for consolidation processes. Instead, it may be necessary to allow participants to rest quietly without any external input.

3.3 Rest-Induced Retrograde Facilitation

Quevedo Pütter, J., & Erdfelder, E. (2024). *Waking rest during retention facilitates memory consolidation, but so does social media use: A storage-retrieval analysis*. Manuscript submitted for publication.

In Manuscript III, we report two laboratory experiments that were designed to compare the effects of *minimal* retroactive interference through waking rest with conditions of both specific and nonspecific retroactive interference. Thus, in comparison with all experiments reported in Manuscripts I and II, these studies may be argued to provide the best conditions for opportunistic consolidation to occur. Moreover, in Experiment 1, we additionally tested memory performances 24 hours after encoding to allow for potential longer-term effects of opportunistic consolidation to emerge.

In both experiments, we aimed to replicate the rest-induced retrograde facilitation effect found by Martini et al. (2020) in relation to post-encoding social media use and to disentangle its underlying mechanisms. To this end, we used the original storage-retrieval MPT model by Riefer and Batchelder (1995) that served as the basis for the adapted model in Manuscript II. This model is tailored to a recognition-then-cued-recall testing procedure for word pairs. Importantly, however, the storage-retrieval analysis relates to the target words only, that is, in the old-new recognition test, participants are asked to decide for a series of *single* words whether they do or do not correspond to the target words previously encoded in association with their respective cue words. Accordingly, storage parameter s , recognition retrieval parameter r_1 , and cued recall retrieval parameter r_2 are assumed to reflect the probabilities of target word storage and retrieval (see Nadarevic, 2017).¹

In Experiment 1, participants learned and immediately recalled a list of 20 Icelandic-German vocabulary pairs before being randomly assigned to one of three experimental conditions. In the waking rest condition, participants were asked to lay their heads

on their arms, close their eyes, and rest quietly for 8 minutes. In the social media condition, participants were asked to use the social media platform Instagram on their own smartphone and from their own account. As a third condition that was not part of the original Martini et al. (2020) study, we included a vocabulary condition in which participants were asked to learn and recall an unrelated list of 20 Norwegian-German vocabulary pairs. Importantly, no German target words from the original learning phase were included in this interpolated learning phase. Next, participants' memory for the Icelandic-German vocabulary was tested in a first delayed recognition-then-cued-recall test sequence. A second delayed recognition-then-cued-recall test sequence took place 24 hours later with a new set of distractor words in the recognition test. Taken together, our experimental design allowed us to compare the effects of nonspecific retroactive interference in the social media condition and specific retroactive interference in the vocabulary condition in relation to conditions of minimal retroactive interference (i.e., the waking rest condition). After the exclusion of outliers, data from $N_1 = 154$ participants was analyzed for the first experimental session and data from $N_2 = 141$ was analyzed for the second session 24 hours later.

With respect to our surface memory measures of cued recall retention and recognition performance immediately after the 8-min retention interval, we observed a memory benefit for the waking rest condition only in comparison to the vocabulary condition but not the social media condition. Thus, we did not replicate the rest-induced retrograde facilitation effect found by Martini et al. (2020) in relation to social media use. Likewise, our model-based analyses revealed no reliable differences in storage or retrieval probabilities between the waking rest and the social media condition. Moreover, these analyses revealed reliably higher storage probabilities (parameter s) in both the waking rest and the social media condition compared to the vocabulary condition. With respect to the cued recall retrieval probabilities (parameter r_2), the same pairwise comparisons revealed reliable differences only for the aggregated data, but not the individual data. For the surface and model-based memory measures after 24 hours, we observed a very similar pattern with a tendency of only slightly decreased descriptive differences.

¹Note that the notation for retrieval parameters r_1 and r_2 in Manuscript III corresponds to the original notation introduced by Riefer and Batchelder (1995). In Manuscript II, retrieval parameter labels were chosen according to the order of memory tests in the free-recall-then-recognition sequence. Thus, parameter r_1 represents recognition retrieval in Manuscript III, whereas it represents free recall retrieval in Manuscript II. Conversely, parameter r_2 represents cued recall retrieval in Manuscript III, whereas it represents free recall retrieval in Manuscript II.

Table 3*Results of the Storage-Retrieval MPT Analysis of Experiment 2 in Manuscript III*

Parameter	Waking rest	Social media	Vocabulary
Aggregated data ^a			
s	.92 [.90, .93]	.92 [.90, .94]	.89 [.87, .91]
r_1	.99 [.99, .99]	.99 [.99, .99]	.99 [.99, .99]
r_2	.63 [.60, .67]	.66 [.63, .69]	.63 [.60, .67]
Individual data ^b			
s	.92 [.90, .94]	.93 [.91, .96]	.89 [.87, .92]
r_1	.99 [.99, .99]	.99 [.99, .99]	.99 [.98, .99]
r_2	.64 [.59, .68]	.66 [.63, .70]	.64 [.59, .69]

Note. Multinomial processing tree (MPT) parameter s = probability of storage of a target word, r_1 = probability of retrieving a target word during recognition, r_2 = probability of retrieving a target word during cued recall. For the estimates of guessing parameter g , please refer to Manuscript II.

^a The model was fitted to the aggregated category frequencies using maximum likelihood (ML) estimation in the R package MPTinR (Singmann & Kellen, 2013). Parameter estimates are presented alongside the corresponding 95% confidence intervals.

^b The model was fitted to the individual category frequencies using Bayesian hierarchical estimation in the R package TreeBUGS (Heck et al., 2018). Posterior means are presented alongside the corresponding 95% Bayesian credibility intervals.

Our results from Experiment 1 suggest (a) that the waking rest versus social media effect found by Martini et al. (2020) might depend on details of the methodological approach that differed between the original and our replication study and (b) that the specific retroactive interference effect in the vocabulary condition was largely driven by storage (i.e., consolidation) processes. Surprisingly, in contrast to our observations in Manuscripts I and II, retrieval processes remained largely unaffected by our manipulation. Moreover, we did not observe any substantial differences between the first and the second delayed memory tests.

In Experiment 2, we set out to confirm the surprising result pattern from Experiment 1. Most importantly, we wanted to test whether our observations would be robust against a reversal of the testing procedure. In Experiment 1, the presentation of target words as part of the recognition test may have served as an undesired relearning opportunity for the subsequent cued recall test. To circumvent this potential problem in Experiment 2, we chose to use a reversed cued-recall-then-recognition testing procedure instead. The corresponding MPT model is mathematically equivalent to the model from Experiment 1. Also, we omitted the second experimental session and focused on short-term effects within a single session in Experiment 2. The final sample size was $N = 157$.

The surface and model-based results confirmed our interim conclusion from Experiment 1. The storage and retrieval MPT parameter estimates are provided in Table 3. We replicated all patterns from Experiment 1 and found no reliable cued recall retrieval (parameter r_2) differences between any conditions for both the aggregated and the individual data. Thus, we found robust evidence against a waking rest versus social media effect across two replication studies and a purely storage-driven effect with respect to the specific retroactive interference induced in the vocabulary condition.

Thereby, the results from both experiments reported in Manuscript III provide the very first direct behavioral evidence in line with the opportunistic consolidation account available in the literature. At the same time, they are at odds with our results from Manuscripts I and II and a passive interference reduction account of forgetting. Given the null-results with respect to differences between the waking rest and the social media condition, minimal retroactive interference does not seem to be a necessary precondition for opportunistic consolidation to unfold. Thus, these results are highly informative for a comprehensive theoretical account of interference-based forgetting that specifies the rather specific conditions under which opportunistic consolidation may be inhibited.

4 Discussion

In each of the three manuscripts included in this thesis, we found evidence for an unambiguous conclusion in line with either the temporal distinctiveness account (Manuscripts I and II) or the opportunistic consolidation account (Manuscript III). Therefore, across the three manuscripts, the general pattern of results is quite complex. In the following, I establish the commonalities and inconsistencies between the three manuscripts before proposing a revised model of interference-based forgetting. After discussing the strengths and limitations of the reported research, I conclude with recommendations for future research.

4.1 Synthesis

Do activities and tasks that reduce, delay, or minimize retroactive interference play a dominant role in everyday forgetting by affecting consolidation processes as suggested by Wixted (2004)? An inspection of only the surface result patterns in all three manuscripts included in this thesis casts doubt on this idea. We observed no significant effects of post-encoding alcohol versus placebo administration on subsequent free or cued recall performance in Manuscript I. In all three experiments of Manuscript II, delaying interpolated learning consistently facilitated subsequent free recall but not recognition. Finally, in both experiments of Manuscript III, post-encoding waking rest facilitated subsequent cued recall and recognition. However, this was only true for the comparison with unrelated vocabulary learning but not with social media use.

This rather mixed result pattern for directly observable recall and recognition performances suggests that retroactive interference effects occur under rather specific conditions, that is, only for specific manipulations and memory tests. Thus, a comprehensive role for consolidation-based retroactive interference is hard to reconcile with our findings, since consolidation effects should have become apparent regardless of specific manipulations (because mere cognitive exertion should be sufficient to inhibit opportunistic consolidation) and memory test choices (because opportunistic consolidation should affect performance even in relatively retrieval-independent memory tests).

The MPT storage-retrieval results from all three manuscripts allow for a more profound interpretation. At first glance, these model-based results seem to diverge between Manuscripts I and II on the one hand, and Manuscript III on the other. The experiments reported in Manuscripts I and II seem to suggest that both alcohol-

induced retrograde facilitation and the temporal gradient of retroactive interference are purely retrieval-driven phenomena without a reliable involvement of storage or consolidation processes. However, such a straightforward interpretation is seriously challenged by our observations from Manuscript III. Here, we found the opposite pattern in both experiments, that is, rest-induced retrograde facilitation in relation to further vocabulary learning was largely based on storage processes, whereas the evidence for a retrieval contribution was unreliable in Experiment 1 and virtually nonexistent in Experiment 2. Overall, this pattern of results implies that both storage and retrieval effects may depend on certain methodological characteristics that differed between our experiments, such as the original and interpolated learning material, the selection and order of memory tests, the duration of the retention interval, the nature of the distractor task during the retention interval, and the applied MPT model.

The absence of reliable storage effects in Manuscripts I and II tentatively suggests that consolidation is affected only if (a) the interpolated learning material is sufficiently similar to the original learning material (as opposed to nonspecific retroactive interference as induced in Manuscript I) and (b) the difference in uninterrupted retroactive interference reduction or minimization between conditions is sufficiently large (as opposed to comparatively small L1-L2 timing differences as induced in Manuscript II). Other factors such as the associative nature of the MPT maintenance parameter in Manuscript I as opposed to the non-associative interpretation of the storage parameters in Manuscripts II and III might also have contributed to the observed divergences (see Hardt et al., 2013; Kuhlmann et al., 2021).

With respect to the role of retrieval processes, the inclusion of multiple retrieval parameters per MPT model allows for a more fine-grained inspection across experiments. Each experiment involved a specific combination of two out of three memory tests: free and cued recall in Manuscript I, free recall and recognition in both MPT experiments in Manuscript II, and cued recall and recognition in both experiments in Manuscript III. Crucially, the retrieval parameter estimates resulting from different models cannot be treated as perfectly equivalent but must be carefully interpreted in the context of the corresponding testing procedure.

Table 4 provides an overview of the retrieval parameter estimates from all MPT experiments in all three manuscripts. As was to be expected by the basic logic of any storage-retrieval model, each model produced ceiling parameter estimates close to 1 for the comparably less retrieval-dependent memory test (i.e., cued recall in Manuscript I, recognition in Manuscripts II and III) and considerably lower parameter estimates for

Table 4*Comparison of MPT Retrieval Parameter Estimates Across All Experiments*

Memory test	Manuscript I		Manuscript II				Manuscript III		
	A	P	High similarity		Low similarity		WR	SM	V
			LS	SL	LS	SL			
Free recall	.50	.44	.64 / .70	.60 / .64	.70	.66	–	–	–
Cued recall	.98	.98	–	–	–	–	.73 / .63	.73 / .66	.68 / .63
Recognition	–	–	.93 / .99	.92 / .99	.99	.99	.98 / .99	.98 / .99	.96 / .99

Note. MPT = multinomial processing tree, A = Alcohol condition, P = Placebo condition, LS = Long-Short L2 timing condition, SL = Short-Long L2 timing condition, WR = Waking rest condition, SM = Social media condition, V = Vocabulary condition. For Experiment 1 in Manuscript III, only estimates from the first session are included. For empty cells, the respective memory measure was not included in the paradigm. In case of two values per cell separated by a slash, the first value refers to the estimate obtained in the respective first MPT experiment (i.e., Experiment 2 in Manuscript II, Experiment 1 in Manuscript III), whereas the second value refers to the estimate obtained in the respective second MPT experiment (i.e., Experiment 3 in Manuscript II, Experiment 2 in Manuscript III). For reasons of simplicity, only the parameter estimates obtained from maximum likelihood estimation for the aggregated data are reported.

the more retrieval-dependent memory test (i.e., free recall in Manuscripts I and II, cued recall in Manuscript III). Interestingly, this pattern implies a severe inconsistency between the cued recall retrieval parameter estimates from Manuscripts I and III: Whereas the probability of cued recall retrieval given successful storage was estimated to be very close to 1 in both the alcohol and the placebo condition in Manuscript I, the same probability was estimated to be substantially lower in all three conditions in Manuscript III.

This inconsistency may be explained by differences in the basic logic of the specific MPT models used in these manuscripts. The free-then-cued-recall model used in Manuscript I is tailored to word pairs as learning material so that its parameters are assumed to be reflective of *associative* memory processing steps: Parameter e denotes the probability of associative encoding of both words of a pair, parameter m the probability of maintaining this association across the retention interval, and parameters r_f and r_c the probabilities of retrieving a successfully stored association during free and cued recall. Thus, it is to be expected that providing the first half of an association (i.e., the first word of the pair), given that this association is available in memory, will almost always lead to successful retrieval of the second half (i.e., the corresponding target word, see Küpper-Tetzel & Erdfelder, 2012).

The recognition-then-cued-recall model as used in Manuscript III is also tailored to word pairs as learning material and was originally intended by Riefer and Batchelder (1995) to reflect associative memory processing steps in the so-called recognition-failure paradigm. However, as pointed out by Nadarevic (2017), it is more reasonable to assume that the model parameters are in fact reflective of *non-associative* processing steps for the target word of each pair. Thus, parameter s denotes the probability of storing (i.e., encoding and maintaining) the target word of a pair, and parameters r_1 and r_2 denote the probabilities of retrieving a successfully stored target word during recognition and cued recall, respectively. The focus on item instead of associative memory of this model is most clearly reflected in the recognition test, as it requires participants to classify single words as “old” or “new” with respect to the previously studied target words without any consideration of the corresponding cue words (see Nadarevic, 2017).

That being said, the non-associative logic of the Riefer and Batchelder (1995) model is not as clear for the cued recall test: Although parameter r_2 is thought to reflect cued recall retrieval for the corresponding target word only, it may actually confound cued recall retrieval and associative storage contributions. Specifically, parameter r_2

reflects a retrieval probability that is conditional on target word storage only. However, for a correct response in a cued recall test, the respective target word must also be stored in association with its corresponding cue word. Put differently, if cued recall fails for a successfully stored target word, it remains unclear whether retrieval was unsuccessful, or whether the target word was simply not stored in association with its corresponding cue word.

These conceptual considerations may explain the inconsistency with respect to cued recall retrieval estimates between Manuscripts I and III. Conceivably, the lower-than-expected cued recall retrieval estimates from Manuscript III represent an unknown combination of associative storage and target retrieval contributions. Given that the probability of target retrieval should be close to 1 given successful target storage, the resulting estimates may be more reflective of associative storage, implying that waking rest facilitated target storage but not associative storage compared to the vocabulary condition. As for target and associative retrieval, this line of argument would imply that the recognition-then-cued-recall paradigm was not ideal for a precise differentiation of retrieval probabilities. Across all three manuscripts, it therefore seems reasonable to suspect that any retroactive interference effects were always at least partially retrieval-driven, provided that the testing procedure involved a free recall test.

To empirically test this post-hoc hypothesis, an additional data set may be considered that I collected as part of an as yet unpublished research project. In an online experiment, specific and nonspecific retroactive interference was manipulated as a between-participants factor with three post-encoding conditions: “easy equations” (low nonspecific and specific retroactive interference), “hard equations” (high nonspecific retroactive interference), and “word pairs” (high specific retroactive interference). The procedure included an immediate cued recall test after the original learning phase followed by a free-then-cued-recall test sequence (after the respective post-encoding activity) for word pairs. This paradigm allowed for an application of the encoding-maintenance-retrieval model by Küpper-Tetzl and Erdfelder (2012). The experiment was preregistered on the Open Science Framework (OSF) as part of a research project involving a series of experiments (osf.io/ykj7b).

In the final sample, $N = 178$ participants ($M_{\text{age}} = 26.64$ years [$SD = 8.81$, $range = 18-62$], $n_{\text{female}} = 118$, $n_{\text{male}} = 53$, $n_{\text{diverse}} = 7$) learned and immediately recalled a list of 40 German word pairs taken from Dimigen et al. (2012). Each word pair was presented for 5 seconds with a 500-ms interstimulus interval. Next, during the 15-min retention

interval, participants performed their randomly assigned post-encoding activity: In the easy equations condition, participants were asked to assess the correctness of very easy mathematical equations as quickly as possible. If they thought that an equation was correct (e.g., $2 + 5 = 7$), they were asked to press the “S” key on their keyboard. If they thought that an equation was incorrect (e.g., $2 + 6 = 7$), they were asked to press the “L” key. In the hard equations condition, participants performed the same task but with considerably more complex correct (e.g., $[13 \cdot 11] + [19 + 43] - [83 - 35] = 157$) and incorrect (e.g., $[12 \cdot 11] + [83 - 26] - [28 + 58] = 102$) mathematical equations. In the word pairs condition, participants were asked to encode and immediately recall two additional lists of 40 word pairs each. None of the additional word pairs had been previously presented as part of the original learning phase. Finally, in the free recall test, participants were given 8 minutes to freely recall as many of the originally learned word pairs as possible, before performing the final cued recall test.

The design of this experiment closely resembles that of both experiments reported in Manuscript III. Specifically, all three experiments included one experimental condition involving very low or minimal specific and nonspecific retroactive interference (i.e., easy equations and waking rest), a second condition involving increased nonspecific retroactive interference (i.e., hard equations and social media), and a third condition involving high specific retroactive interference (i.e., word pairs and vocabulary). As the additionally reported experiment included a final free recall, it allows for an indirect test of the post-hoc explanation for our null-findings with respect to cued recall retrieval in Manuscript III.

The most parsimonious version of the encoding-maintenance-retrieval model did not fit the data well, neither for the individual data according to posterior-predictive p -values (see Heck et al., 2018), $p_1 < .001$, $p_2 = .108$ in the easy equations condition, $p_1 = .041$, $p_2 = .388$ in the hard equations condition, $p_1 = .193$, $p_2 = .082$ in the word pairs condition, nor for the aggregated data, $G^2(15) = 80.87$, $p < .001$.

Several less restrictive model versions were inspected to find a model version with acceptable fit to the data. A generalized model version was specified with two free recall retrieval parameters r_f (r_{fs} following successful immediate cued recall, r_{fu} following unsuccessful immediate cued recall), two cued recall retrieval parameters r_c (r_{cs} for the final cued recall following successful free recall, r_{cu} for the immediate cued recall and for the final cued recall following unsuccessful free recall), and two single word retrieval parameters u given unsuccessful encoding or maintenance (u_s following successful cued recall, u_u following unsuccessful cued recall).

Table 5*Results of an Additional Encoding-Maintenance-Retrieval MPT Analysis*

Parameter	Easy equations	Hard equations	Word pairs
e	.58 [.53, .63]	.54 [.49, .59]	.58 [.54, .61]
m	.93 [.90, .96]	.93 [.90, .96]	.87 [.83, .92]
r_{fs}	.57 [.53, .61]	.59 [.54, .63]	.30 [.25, .36]
r_{fu}	.22 [.06, .38]	.18 [.01, .52]	.18 [.05, .34]
r_{cs}	.98 [.96, .99]	.98 [.96, .99]	.97 [.94, .99]
r_{cu}	.94 [.92, .96]	.96 [.94, .98]	.95 [.93, .97]

Note. Multinomial processing tree (MPT) parameter e = probability of associative encoding of a word pair, m = probability of associative maintenance of a word pair across the retention interval, r_{fs} = probability of retrieving a word pair as an association during the final free recall following successful immediate cued recall, r_{fu} = probability of retrieving a word pair as an association during the final free recall following unsuccessful immediate cued recall, r_{cs} = probability of retrieving a word pair as an association during the final cued recall following successful free recall, r_{cu} = probability of retrieving a word pair as an association during the immediate cued recall and during the final cued recall following unsuccessful free recall. The model was fitted to the individual category frequencies using Bayesian hierarchical estimation in the R package TreeBUGS (Heck et al., 2018). Posterior means are presented alongside the corresponding 95% Bayesian credibility intervals.

This generalized model version fit the individual data well, $p_1 = .084$, $p_2 = .377$ in the easy equations condition, $p_1 = .477$, $p_2 = .439$ in the hard equations condition, $p_1 = .451$, $p_2 = .414$ in the word pairs condition. However, there was still considerable misfit for the aggregated data, $G^2(6) = 28.38$, $p < .001$. Since no conceivable model version fit the aggregated data, further MPT analyses were only conducted based on the individual data. Corresponding parameter estimates are provided in Table 5.

Overall, maintenance and free recall retrieval parameter estimates were not reliably higher in the easy equations compared to the hard equations condition, that is, all Bayesian $p > .05$. In contrast, we did observe reliably higher estimates in the easy equations compared to the word pairs condition for maintenance parameter m , Bayesian

$p = .018$, and for free recall retrieval parameter r_{fs} , Bayesian $p < .001$. The same held true for the comparison between the hard equations and the word pairs condition, Bayesian $p = .018$ for parameter m , Bayesian $p < .001$ for parameter r_{fs} . The data and analysis scripts necessary to reproduce the reported results are publicly available on the OSF (osf.io/h8w9z/?view_only=c6ac92e3563a48338a497cebd533964c).

These results provide tentative evidence in favor of the post-hoc explanation with respect to our null-findings for cued recall retrieval in Manuscript III. Given the similarity between the experimental conditions used in the additional experiment and those reported in Manuscript III, it seems reasonable to suspect that we would have observed a reliable retrieval effect in Manuscript III had we adapted the paradigm to include a free recall test.

Accepting this explanation, the general conclusion across all three manuscripts of this thesis may be summarized in one central statement: Retroactive interference *always* affects retrieval, but only *sometimes* consolidation. This conclusion is at odds with both the diversion-similarity retroactive interference account by Dewar et al. (2007) and the temporal distinctiveness account by Brown et al. (2007), and necessitates a revised model of interference-based forgetting.

4.2 A Revised Model of Interference-Based Forgetting

The observation that storage processes were only affected by the specific retroactive interference induced in the vocabulary condition in Manuscript III suggests that a rather high degree of similarity between original and interpolated learning is a necessary precondition for opportunistic consolidation to be inhibited during wakefulness. It seems highly unlikely that the sleep-induced maintenance benefit observed by Erdfelder et al. (2022) was the result of participants in the wake condition engaging in sufficiently similar learning activities during the 12-hr retention interval. Instead, this finding is more easily explained in terms of unique-to-sleep consolidation. In other words, whereas sleep-specific theories such as the active systems consolidation model (Rasch & Born, 2013) may account for forgetting processes during sleep, models of interference-based forgetting during wakefulness need to be adjusted in light of the findings presented in this thesis.

To this end, the distinction of opportunistic consolidation and passive interference reduction proposed by Mednick et al. (2011) may serve as a starting point. In line with passive interference reduction accounts such as the temporal distinctiveness account, we observed consistent retrieval effects in Manuscripts I and II, and the absence of

such effects in Manuscript III may well be explained in terms of the test procedure (see Synthesis section). Crucially, retrieval was not only hampered by the similar interpolated materials used in Experiment 2 and in the high similarity condition of Experiment 3 in Manuscript II, but also by nonspecific activities such as watching movies in the placebo condition in Manuscript I and by studying geometric figures in the low similarity condition of Experiment 3 in Manuscript II. Thus, the distinction of specific and nonspecific retroactive interference seems to be negligible with respect to retrieval processes. Instead, any kind of task or material may interfere with subsequent retrieval.

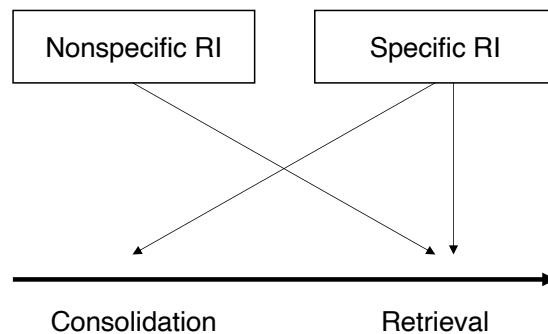
Interestingly, post-encoding social media use in Manuscript III did not interfere with retrieval compared to an equally long period of waking rest. This finding may be interpreted in different ways. First, it may be the case that participants in the waking rest condition engaged in mental activities such as mind-wandering or autobiographical thinking that amounted to a degree of nonspecific retroactive interference comparable to that induced by social media (see Varma et al., 2017). Second, it may be assumed that some minimum degree of nonspecific or specific retroactive interference needs to be induced either in terms of duration (7 hours of nonspecific retroactive interference in Manuscript I versus only 8 minutes in the social media condition in Manuscript III) or in terms of encoding strength (intentional learning in Manuscript II and in the vocabulary condition in Manuscript III versus incidental learning in the social media condition in Manuscript III) to affect retrieval.

The general observation of consistent retrieval effects is in line with a temporal distinctiveness account and, more generally, any passive interference reduction account. Whereas it might be argued that the assumptions of the temporal distinctiveness account are overly specific with respect to alcohol- and rest-induced retrograde facilitation, the temporal gradient of retroactive interference cannot be explained without the temporal specifications of this account. Thus, the temporal distinctiveness account offers an appropriate explanation for retrieval effects in interference-based forgetting. That being said, temporal distinctiveness cannot account for the storage effects observed in Manuscript III. Instead, this aspect of the result pattern necessitates a more integrative theoretical account.

To explain the storage effects observed in Manuscript III, the opportunistic consolidation account proposed by Mednick et al. (2011) needs to be revised. One key assumption of the original model is that encoding and consolidation compete for one shared pool of hippocampal resources and are therefore mutually exclusive.

Figure 3

Proposal for a Revised Model of Interference-Based Forgetting



Note. Nonspecific retroactive interference (RI) is induced by any kind of interpolated cognitive processing and encoding demands, whereas specific RI is only induced by interpolated learning material similar to the original learning material.

However, the absence of storage effects in Manuscripts I and II suggests that these resources may be more specific than originally assumed. Instead, the hippocampus may be able to simultaneously encode new information and consolidate older memories, provided that original and new pieces of information are sufficiently dissimilar and therefore depend on non-overlapping networks within the hippocampus. It follows that the interpolated presentation of Norwegian-German word pairs in Manuscript III inhibited consolidation not because *any* encoding occurred but because the word pairs were sufficiently similar to the original Icelandic-German word pairs to usurp shared hippocampal resources. In contrast, the activities performed by participants in the social media condition in Manuscript III and also in the placebo condition in Manuscript I did not interfere with consolidation processes because they may have involved different hippocampal networks.

That being said, even such a revised version of the opportunistic consolidation account is difficult to reconcile with the observation of no storage effects in Manuscript II. In both Experiment 2 and 3, the MPT storage parameter remained unaffected by the supposedly specific retroactive interference induced by interpolated word lists that were generated from the same item pool as the original word lists. As a tentative auxiliary assumption, it may be argued that consolidation-based retroactive interference does not follow a temporal gradient, at least during the first minutes after encoding. Instead, in Manuscript II, consolidation of L1 items may simply have

resumed after the end of the respective L2 learning phase, resulting in equal degrees of consolidation regardless of the temporal position of interpolated L2 learning (see Ecker et al., 2015). Thus, it may be necessary to manipulate the overall duration of interpolated learning and not merely its temporal position to observe storage effects.

Taken together, based on the findings from this thesis, I tentatively propose an integrative model of interference-based forgetting that combines the temporal distinctiveness account by Brown et al. (2007) with elements of the opportunistic consolidation account by Mednick et al. (2011). This model assumes that the relationship between specific and nonspecific retroactive interference on the one hand and consolidation and retrieval on the other is exactly reversed compared to the diversion-similarity retroactive interference account by Dewar et al. (2007). Thus, both nonspecific and specific retroactive interference affect retrieval, but only specific retroactive interference may additionally interfere with consolidation (see Figure 3). Such a model not only explains the general pattern observed in the research reported in this thesis, but may also be used to derive new predictions for future research (see Future Research section).

4.3 Strengths and Limitations

The research conducted as part of this thesis advances our understanding of interference-based forgetting in important ways. At the same time, certain limitations should be acknowledged. In the following, both strengths and limitations of the reported research are reflected.

One major strength of this thesis lies in the introduction of storage-retrieval MPT modeling into the investigation of opportunistic consolidation and temporal distinctiveness. Thereby, the presumably selective effects of specific and nonspecific retroactive interference on consolidation and retrieval processes become empirically testable. The advantage of storage-retrieval MPT analyses compared to conventional analyses of directly observable memory measures such as recall and recognition may most clearly be illustrated with respect to the diversion-similarity retroactive interference model by Dewar et al. (2007): Whereas conventional memory measures at best allow for very indirect partial tests of this account (see, e.g., the experiment reported by Dewar et al., 2007), an appropriate storage-retrieval MPT model allows to precisely disentangle storage (i.e., consolidation) and retrieval contributions to memory performance following specific versus nonspecific (i.e., similarity versus diversion) retroactive interference.

This thesis also demonstrates the flexibility of storage-retrieval MPT models: Across the three manuscripts, well-established models were either used in their original form (see Manuscripts I and III), or they were carefully adapted to a specific to-be-replicated paradigm from the literature (see Manuscripts II and III). Thus, researchers are not bound to specific models or paradigms but may instead adapt both to match their research questions.

At the same time, such flexibility in the application of MPT models also means that some of the models applied in this thesis have thus far not been subjected to rigorous validation studies. Importantly, the construct validity of model parameters and the corresponding latent processes of newly developed MPT models needs to be evaluated (Schmidt et al., 2023). Ideally, the selective influence of experimental manipulations on the model parameters should be assessed before using a new model to answer substantive research questions. For example, for the encoding-maintenance-retrieval model by Küpper-Tetzel and Erdfelder (2012) used in Manuscript I, it has been shown that the encoding parameter e is not affected by the length of the retention interval, that both the maintenance parameter m and the free recall retrieval parameter r_f are negatively affected by longer retention intervals, and that only r_f (but neither e nor m) is positively affected by the presentation of retrieval cues during free recall (Erdfelder et al., 2022; Küpper-Tetzel & Erdfelder, 2012). Likewise, the storage-retrieval model used in Experiment 1 in Manuscript III has been rigorously validated by Riefer and Batchelder (1995) by assessing selective effects of a wide range of experimental manipulations on storage parameter s , recognition retrieval parameter r_1 , and cued recall retrieval parameter r_2 .

Although it might be argued that the validation results reported by Riefer and Batchelder (1995) also apply to some degree to the adapted model versions used in Experiments 2 and 3 in Manuscript II (tailored to a free-recall-then-recognition procedure for single words) and in Experiment 2 in Manuscript III (applied to a reversed cued-recall-then-recognition procedure), these model versions have thus far not been explicitly validated. However, some evidence in favor of both models may be derived from our results. First, across both Experiments 2 and 3 in Manuscript II, we observed a selective influence of our L2 timing manipulation on free recall retrieval parameter r_1 . Such a selective effect is predicted by the temporal distinctiveness account and implies that the model allows for a clear-cut differentiation of storage and retrieval contributions to free recall performance. At the same time, the recognition parameter r_2 was estimated to be close to 1 in both experiments. This finding is in

line with the expectation of nearly perfect recognition when the corresponding word was successfully stored in memory.

Second, the adapted model version used in Experiment 2 in Manuscript III is actually mathematically equivalent to the original model. Given that parameter estimates in Experiment 2 were quite similar to those obtained with the original paradigm in Experiment 1, it seems reasonable to expect the validation results by Riefer and Batchelder (1995) to generalize to a reversed memory test procedure. Again, the recognition retrieval parameter r_1 was estimated to be close to 1 in both experiments. Also, we observed a selective influence of our experimental manipulation on storage parameter s particularly in Experiment 2 but also based on the individual data in Experiment 1. Again, such an observation implies that the model indeed allows for a differentiation of storage and retrieval contributions.

Another limitation of the present research may be the possibility of intentional rehearsal and other memory-relevant activities during the retention interval. Indeed, this represents a challenge for any study on opportunistic consolidation since largely uncontrolled retention intervals are typically regarded as a necessary and even desirable feature of waking rest studies (see Wixted, 2004). Thus, it may be the case that reduced, delayed, and minimized retroactive interference (i.e., alcohol condition in Manuscript I, LS condition in Manuscript II, waking rest condition in Manuscript III) is confounded with an increased possibility for intentional rehearsal of the previously encoded material.

In the literature, the use of difficult-to-rehearse materials (e.g., non-words, Dewar et al., 2014) and the collection of self-reports (e.g., Martini et al., 2020) have been proposed as possible remedies. Self-reported rehearsal during the retention interval was assessed in all three manuscripts included in this thesis. To ensure the robustness of our conclusions, we conducted sensitivity analyses without the data of participants who reported to have engaged in intentional rehearsal. Given the ubiquity of rehearsal in Experiment 3 in Manuscript II, we also included participants' rehearsal ratings in an ANCOVA to make sure that there was no significant interaction effect of rehearsal and our experimental manipulation on free recall performance. Moreover, our storage-retrieval MPT modeling approach allows for an additional check for any unwanted influence of rehearsal: Differences in rehearsal between conditions should arguably result in differences in both storage (or maintenance) and retrieval parameters. However, with the exception of Experiment 1 in Manuscript III, we only observed selective effects of our manipulations on either storage or retrieval processes. Thus, it

seems unlikely that intentional rehearsal may have significantly contributed to our result patterns.

Finally, the manuscripts included in this thesis stand out from most of the literature on opportunistic consolidation and temporal distinctiveness due to the comprehensive application of open science practices, particularly the registered report format of Manuscript I and the detailed preregistration of all experiments reported in Manuscripts II and III. Thereby, our hypothesis tests were made fully transparent and reproducible (Lakens, 2019). In contrast to rather small sample sizes reported for at least some studies in the waking rest literature (see Humiston et al., 2019), our sample size rationale involved rigorous a priori power analyses for all experiments. We also applied innovative analysis tools such as a sequential testing procedure (i.e., the sequential probability ratio t test, see Schnuerch & Erdfelder, 2020) in Manuscript I and an interaction criterion for replication success (see Anderson & Maxwell, 2016) in Experiment 1 in Manuscript II. Moreover, we conducted internal replications in Manuscripts II and III, and multiverse or sensitivity analyses in all manuscripts. Taken together, these measures justify a particularly high confidence in the results reported in this thesis.

4.4 Future Research

Future research may build on the conclusions from this thesis by further specifying the exact conditions under which unique-to-sleep consolidation, opportunistic consolidation, and passive interference reduction contribute to interference-based forgetting. The revised model proposed above may serve as a theoretical basis for future research. Key predictions to be derived from this model are the following:

1. When retroactive interference is observed in free recall, the effect should always be driven by retrieval, never by consolidation alone.
2. Consolidation should additionally contribute to retroactive interference only if the original and interpolated learning materials are relatively similar (i.e., if their encoding and consolidation relies on the same pool of hippocampal resources) and the difference in retroactive interference between conditions is sufficiently large.

To empirically test these predictions, conventional memory measures are insufficient. Instead, researchers may adopt the storage-retrieval MPT modeling approach as

demonstrated in this thesis. New models may be developed and validated to extend the scope of possible applications to previously used paradigms.

Future research may focus on specifying the conditions under which opportunistic consolidation contributes to retroactive interference effects. The revised model tentatively suggests that consolidation during wakefulness is only affected by specific retroactive interference. In other words, increased similarity between original and interpolated learning should decrease the probability of successful consolidation. This prediction may be rigorously tested by systematically manipulating the similarity of original and interpolated learning materials. For example, the classic study by McGeoch and McDonald (1931, see Introduction section) may be conceptually replicated and the resulting data analyzed by means of an appropriate storage-retrieval MPT model. From the revised model, it follows that both storage and retrieval probabilities should decrease with increased similarity.

Likewise, the strength of nonspecific retroactive interference may be manipulated to scrutinize its presumably selective influence on retrieval probabilities. For this purpose, minimal retroactive interference may be induced by means of a waking rest condition. In contrast to the social media condition used in Manuscript III, further experimental conditions may be devised that allow for more experimental control. For example, the d2 test of attention (see, e.g., Marhenke et al., 2023) may be adapted to induce different degrees of cognitive demands (e.g., by manipulating time pressure or the similarity of target and distractor items). From the revised model, it follows that only retrieval but not storage (i.e., consolidation) should be affected by increased cognitive demands.

The conditions under which opportunistic consolidation may or may not contribute to the temporal gradient of retroactive interference seem to be most difficult to derive from the revised model. As it may be assumed that consolidation simply resumed after the end of the respective L2 learning phase in Experiments 2 and 3 in Manuscript II (see A Revised Model of Interference-Based Forgetting section), future research may replicate this basic paradigm with longer and more cognitively demanding learning phases that may more effectively disrupt the consolidation process.

As for sleep-induced retrograde facilitation, future research may more explicitly consider the activities performed by participants in the wake condition during the retention interval. The revised model suggests that increasing nonspecific retroactive interference during wakefulness should increase retrieval contributions, whereas increasing specific retroactive interference (e.g., by having participants study additional

materials during the retention interval) should increase the contributions of both retrieval and opportunistic consolidation processes.

On a more general note, future research may more explicitly consider intentional rehearsal during the retention interval, for example by using incidental rather than intentional learning instructions. Moreover, to account for the distinction between unique-to-sleep and opportunistic consolidation, the influence of sleep during retention intervals that last several days should be more explicitly considered. A research program along these lines may ultimately result in a more precise model of interference-based forgetting during sleep and wakefulness that integrates the temporal distinctiveness account with a well-specified consolidation mechanism.

4.5 Conclusion

Memory research from the past decades has ascribed consolidation a comprehensive role in interference-based forgetting. Indeed, the results from this thesis do not imply that the inhibition versus facilitation of consolidation is irrelevant for forgetting. Quite to the contrary, this thesis suggests that unique-to-sleep consolidation is a major contributor to sleep-induced retrograde facilitation. However, opportunistic consolidation seems to be far less relevant for interference-based forgetting than previously assumed. Instead, retrieval benefits from passive interference reduction seem to explain most phenomena of interference-based forgetting during wakefulness. Whereas the exact conditions under which opportunistic consolidation might play an additional role need to be further determined by future research, this thesis ascribes temporal distinctiveness a central role in daytime forgetting. By advancing theoretical accounts of interference-based forgetting, it effectively addresses the “laundry list” critique by Wixted (2004).

5 Bibliography

- Anderson, S. F., & Maxwell, S. E. (2016). There's more than one way to conduct a replication study: Beyond statistical significance. *Psychological Methods*, *21*(1), 1–12. <https://doi.org/10.1037/met0000051>
- Antony, J. W., Romero, A., Vierra, A. H., Luenser, R. S., Hawkins, R. D., & Bennion, K. A. (2022). Semantic relatedness retroactively boosts memory and promotes memory interdependence across episodes. *eLife*, *11*, Article e72519. <https://doi.org/10.7554/eLife.72519>
- Batchelder, W. H., & Riefer, D. M. (1980). Separation of storage and retrieval factors in free recall of clusterable pairs. *Psychological Review*, *87*(4), 375–397. <https://doi.org/10.1037/0033-295X.87.4.375>
- Batchelder, W. H., & Riefer, D. M. (1999). Theoretical and empirical review of multinomial process tree modeling. *Psychonomic Bulletin & Review*, *6*(1), 57–86. <https://doi.org/10.3758/BF03210812>
- Bäuml, K.-H. (1991a). Experimental analysis of storage and retrieval processes involved in retroactive inhibition: The effect of presentation mode. *Acta Psychologica*, *77*(2), 103–119. [https://doi.org/10.1016/0001-6918\(91\)90026-V](https://doi.org/10.1016/0001-6918(91)90026-V)
- Bäuml, K.-H. (1991b). Retroaktive Hemmung: Der Einfluß des interpolierten Kategorienmaterials auf die Verfügbarkeit von Information [Retroactive inhibition: The influence of interpolated category material on the availability of information]. *Zeitschrift für Experimentelle und Angewandte Psychologie*, *38*(2), 169–187. <https://doi.org/10.5283/epub.9027>
- Bäuml, K.-H. (1991c). Retroaktive Hemmung: Kein Einfluss der Kategorisierungsart—konzeptuell versus phonetisch-visuell—auf die Verfügbarkeit von Informationen [Retroactive inhibition: No influence of the type of categorization—conceptual versus phonetic-visual—on the availability of information]. *Zeitschrift für Psychologie*, *199*, 177–190. <https://doi.org/10.5283/epub.9024>
- Berres, S. (2023). *The sleep benefit in episodic memory: Investigating underlying mechanisms* [Doctoral dissertation, University of Mannheim]. Mannheim Electronic Document Server (MADOC). <https://madoc.bib.uni-mannheim.de/64706>
- Berres, S., & Erdfelder, E. (2021). The sleep benefit in episodic memory: An integrative review and a meta-analysis. *Psychological Bulletin*, *147*(12), 1309–1353. <https://doi.org/10.1037/bul0000350>

- Bower, G. H., Thompson-Schill, S., & Tulving, E. (1994). Reducing retroactive interference: An interference analysis. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *20*(1), 51–66. <https://doi.org/10.1037/0278-7393.20.1.51>
- Brown, G. D. A., Neath, I., & Chater, N. (2007). A temporal ratio model of memory. *Psychological Review*, *114*(3), 539–576. <https://doi.org/10.1037/0033-295X.114.3.539>
- Bruce, K. R., & Pihl, R. O. (1997). Forget "drinking to forget": Enhanced consolidation of emotionally charged memory by alcohol. *Experimental and Clinical Psychopharmacology*, *5*(3), 242–250. <https://doi.org/10.1037/1064-1297.5.3.242>
- Carlyle, M., Dumay, N., Roberts, K., McAndrew, A., Stevens, T., Lawn, W., & Morgan, C. J. A. (2017). Improved memory for information learnt before alcohol use in social drinkers tested in a naturalistic setting. *Scientific Reports*, *7*, Article 6213. <https://doi.org/10.1038/s41598-017-06305-w>
- Dewar, M., Alber, J., Butler, C., Cowan, N., & Della Sala, S. (2012). Brief wakeful resting boosts new memories over the long term. *Psychological Science*, *23*(9), 955–960. <https://doi.org/10.1177/0956797612441220>
- Dewar, M., Alber, J., Cowan, N., & Della Sala, S. (2014). Boosting long-term memory via wakeful rest: Intentional rehearsal is not necessary, consolidation is sufficient. *PLOS ONE*, *9*(10), Article e109542. <https://doi.org/10.1371/journal.pone.0109542>
- Dewar, M., Cowan, N., & Della Sala, S. (2007). Forgetting due to retroactive interference: A fusion of Müller and Pilzecker's (1900) early insights into everyday forgetting and recent research on anterograde amnesia. *Cortex*, *43*(5), 616–634. [https://doi.org/10.1016/S0010-9452\(08\)70492-1](https://doi.org/10.1016/S0010-9452(08)70492-1)
- Dimigen, O., Kliegl, R., & Sommer, W. (2012). Trans-saccadic parafoveal preview benefits in fluent reading: A study with fixation-related brain potentials. *NeuroImage*, *62*(1), 381–393. <https://doi.org/10.1016/j.neuroimage.2012.04.006>
- Drachman, D. A., & Leavitt, J. (1972). Memory impairment in the aged: Storage versus retrieval deficit. *Journal of Experimental Psychology*, *93*(2), 302–308. <https://doi.org/10.1037/h0032489>
- Dudai, Y. (2004). The neurobiology of consolidations, or, how stable is the engram? *Annual Review of Psychology*, *55*, 51–86. <https://doi.org/10.1146/annurev.psych.55.090902.142050>
- Ebbinghaus, H. (1885). *Über das Gedächtnis: Untersuchungen zur experimentellen Psychologie* [About memory: Studies in experimental psychology]. Duncker & Humblot.

- Ecker, U. K. H., Brown, G. D. A., & Lewandowsky, S. (2015). Memory without consolidation: Temporal distinctiveness explains retroactive interference. *Cognitive Science*, *39*(7), 1570–1593. <https://doi.org/10.1111/cogs.12214>
- Erdfelder, E., Auer, T.-S., Hilbig, B. E., Aßfalg, A., Moshagen, M., & Nadarevic, L. (2009). Multinomial processing tree models: A review of the literature. *Zeitschrift für Psychologie / Journal of Psychology*, *217*(3), 108–124. <https://doi.org/10.1027/0044-3409.217.3.108>
- Erdfelder, E., Berres, S., Quevedo Pütter, J., & Küpper-Tetzl, C. E. (2022). *Why does sleep improve episodic memory? An encoding-maintenance-retrieval analysis*. Manuscript under revision.
- Erdfelder, E., Quevedo Pütter, J., & Schnuerch, M. (in press). On aggregation invariance of multinomial processing tree models. *Behavior Research Methods*.
- Hardt, O., Nader, K., & Nadel, L. (2013). Decay happens: The role of active forgetting in memory. *Trends in Cognitive Sciences*, *17*(3), 111–120. <https://doi.org/10.1016/j.tics.2013.01.001>
- Heck, D. W., Arnold, N. R., & Arnold, D. (2018). TreeBUGS: An R package for hierarchical multinomial-processing-tree modeling. *Behavior Research Methods*, *50*(1), 264–284. <https://doi.org/10.3758/s13428-017-0869-7>
- Hogan, R. M., & Kintsch, W. (1971). Differential effects of study and test trials on long-term recognition and recall. *Journal of Verbal Learning and Verbal Behavior*, *10*(5), 562–567. [https://doi.org/10.1016/S0022-5371\(71\)80029-4](https://doi.org/10.1016/S0022-5371(71)80029-4)
- Humiston, G. B., Tucker, M. A., Summer, T., & Wamsley, E. J. (2019). Resting states and memory consolidation: A preregistered replication and meta-analysis. *Scientific Reports*, *9*, Article 19345. <https://doi.org/10.1038/s41598-019-56033-6>
- Jenkins, J. G., & Dallenbach, K. M. (1924). Obliviscence during sleep and waking. *The American Journal of Psychology*, *35*(4), 605–612. <https://doi.org/10.2307/1414040>
- Klauer, K. C. (2010). Hierarchical multinomial processing tree models: A latent-trait approach. *Psychometrika*, *75*(1), 70–98. <https://doi.org/10.1007/s11336-009-9141-0>
- Knowles, S. K. Z., & Duka, T. (2004). Does alcohol affect memory for emotional and non-emotional experiences in different ways? *Behavioural Pharmacology*, *15*(2), 111–121. <https://doi.org/10.1097/00008877-200403000-00003>

- Kuhlmann, B. G., Brubaker, M. S., Pfeiffer, T., & Naveh-Benjamin, M. (2021). Longer resistance of associative versus item memory to interference-based forgetting, even in older adults. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *47*(3), 422–438. <https://doi.org/10.1037/xlm0000963>
- Küpper-Tetzl, C. E., & Erdfelder, E. (2012). Encoding, maintenance, and retrieval processes in the lag effect: A multinomial processing tree analysis. *Memory*, *20*(1), 37–47. <https://doi.org/10.1080/09658211.2011.631550>
- Lakens, D. (2019). The value of preregistration for psychological science: A conceptual analysis. *Japanese Psychological Review*, *62*(3), 221–230. https://doi.org/10.24602/sjpr.62.3_221
- Lamberty, G. J., Beckwith, B. E., Petros, T. V., & Ross, A. R. (1990). Posttrial treatment with ethanol enhances recall of prose narratives. *Physiology & Behavior*, *48*(5), 653–658. [https://doi.org/10.1016/0031-9384\(90\)90206-J](https://doi.org/10.1016/0031-9384(90)90206-J)
- Mann, R. E., Cho-Young, J., & Vogel-Sprott, M. (1984). Retrograde enhancement by alcohol of delayed free recall performance. *Pharmacology Biochemistry and Behavior*, *20*(4), 639–642. [https://doi.org/10.1016/0091-3057\(84\)90317-4](https://doi.org/10.1016/0091-3057(84)90317-4)
- Marhenke, R., Acevedo, B., Sachse, P., & Martini, M. (2023). Individual differences in sensory processing sensitivity amplify effects of post-learning activity for better and for worse. *Scientific Reports*, *13*(1), 4451. <https://doi.org/10.1038/s41598-023-31192-9>
- Martini, M., Heinz, A., Hinterholzer, J., Martini, C., & Sachse, P. (2020). Effects of wakeful resting versus social media usage after learning on the retention of new memories. *Applied Cognitive Psychology*, *34*(2), 551–558. <https://doi.org/10.1002/acp.3641>
- Martini, M., & Sachse, P. (2020). Factors modulating the effects of waking rest on memory. *Cognitive Processing*, *21*(1), 149–153. <https://doi.org/10.1007/s10339-019-00942-x>
- McGaugh, J. L. (2000). Memory—a century of consolidation. *Science*, *287*(5451), 248–251. <https://doi.org/10.1126/science.287.5451.248>
- McGeoch, J. A., & McDonald, W. T. (1931). Meaningful relation and retroactive inhibition. *The American Journal of Psychology*, *43*(4), 579–588. <https://doi.org/10.2307/1415159>
- Mednick, S. C., Cai, D. J., Shuman, T., Anagnostaras, S., & Wixted, J. T. (2011). An opportunistic theory of cellular and systems consolidation. *Trends in Neurosciences*, *34*(10), 504–514. <https://doi.org/10.1016/j.tins.2011.06.003>

- Mercer, T. (2015). Wakeful rest alleviates interference-based forgetting. *Memory*, *23*(2), 127–137. <https://doi.org/10.1080/09658211.2013.872279>
- Moshagen, M. (2010). multiTree: A computer program for the analysis of multinomial processing tree models. *Behavior Research Methods*, *42*(1), 42–54. <https://doi.org/10.3758/BRM.42.1.42>
- Mueller, C. W., Lisman, S. A., & Spear, N. E. (1983). Alcohol enhancement of human memory: Tests of consolidation and interference hypotheses. *Psychopharmacology*, *80*(3), 226–230. <https://doi.org/10.1007/BF00436158>
- Müller, G. E., & Pilzecker, A. (1900). Experimentelle Beiträge zur Lehre vom Gedächtniss [Experimental contributions to the science of memory]. *Zeitschrift für Psychologie, Ergänzungsband [Supplement] 1*, 1–300.
- Nadarevic, L. (2017). Emotionally enhanced memory for negatively arousing words: Storage or retrieval advantage? *Cognition & Emotion*, *31*(8), 1557–1570. <https://doi.org/10.1080/02699931.2016.1242477>
- Nørby, S. (2015). Why forget? On the adaptive value of memory loss. *Perspectives on Psychological Science*, *10*(5), 551–578. <https://doi.org/10.1177/1745691615596787>
- Parker, E. S., Birnbaum, I. M., Weingartner, H., Hartley, J. T., Stillman, R. C., & Wyatt, R. J. (1980). Retrograde enhancement of human memory with alcohol. *Psychopharmacology*, *69*(2), 219–222. <https://doi.org/10.1007/BF00427653>
- Parker, E. S., Morihisa, J. M., Wyatt, R. J., Schwartz, B. L., Weingartner, H., & Stillman, R. C. (1981). The alcohol facilitation effect on memory: A dose-response study. *Psychopharmacology*, *74*(1), 88–92. <https://doi.org/10.1007/BF00431763>
- Quevedo Pütter, J., Dahler, S., & Erdfelder, E. (2024). *Opportunistic consolidation or temporal distinctiveness? Retrieval, not storage, drives the temporal gradient of retroactive interference in episodic memory*. Manuscript submitted for publication.
- Quevedo Pütter, J., & Erdfelder, E. (2022). Alcohol-Induced Retrograde Facilitation? *Experimental Psychology*, *69*(6), 335–350. <https://doi.org/10.1027/1618-3169/a000569>
- Quevedo Pütter, J., & Erdfelder, E. (2024). *Waking rest during retention facilitates memory consolidation, but so does social media use: A storage-retrieval analysis*. Manuscript submitted for publication.
- Rasch, B., & Born, J. (2013). About sleep’s role in memory. *Physiological Reviews*, *93*(2), 681–766. <https://doi.org/10.1152/physrev.00032.2012>

- Riefer, D. M., & Batchelder, W. H. (1988). Multinomial modeling and the measurement of cognitive processes. *Psychological Review*, *95*(3), 318–339. <https://doi.org/10.1037/0033-295X.95.3.318>
- Riefer, D. M., & Batchelder, W. H. (1995). A multinomial modeling analysis of the recognition-failure paradigm. *Memory & Cognition*, *23*(5), 611–630. <https://doi.org/10.3758/BF03197263>
- Rouder, J. N., & Batchelder, W. H. (1998). Multinomial models for measuring storage and retrieval processes in paired associate learning. In C. E. Dowling, F. S. Roberts, & P. Theuns (Eds.), *Recent Progress in Mathematical Psychology* (pp. 195–226). Psychology Press.
- Sayette, M. A., Reichle, E. D., & Schooler, J. W. (2009). Lost in the sauce: The effects of alcohol on mind wandering. *Psychological Science*, *20*(6), 747–752. <https://doi.org/10.1111/j.1467-9280.2009.02351.x>
- Schmidt, O., Erdfelder, E., & Heck, D. W. (2023). How to develop, test, and extend multinomial processing tree models: A tutorial. *Psychological Methods*. Advance online publication. <https://doi.org/10.1037/met0000561>
- Schnuerch, M., & Erdfelder, E. (2020). Controlling decision errors with minimal costs: The sequential probability ratio t test. *Psychological Methods*, *25*(2), 206–226. <https://doi.org/10.1037/met0000234>
- Singmann, H., Heck, D. W., Barth, M., Erdfelder, E., Arnold, N. R., Aust, F., Calanchini, J., Gümüşdaglı, F. E., Horn, S. S., Kellen, D., Klauer, K. C., Matzke, D., Meissner, F., Michalkiewicz, M., Schaper, M. L., Stahl, C., Kuhlmann, B. G., & Groß, J. (2024). Evaluating the robustness of parameter estimates in cognitive models: A meta-analytic review of multinomial processing tree models across the multiverse of estimation methods. *Psychological Bulletin*. Advance online publication. <https://doi.org/10.1037/bul0000434>
- Singmann, H., & Kellen, D. (2013). MPTinR: Analysis of multinomial processing tree models in R. *Behavior Research Methods*, *45*(2), 560–575. <https://doi.org/10.3758/s13428-012-0259-0>
- Smith, J. B., & Batchelder, W. H. (2008). Assessing individual differences in categorical data. *Psychonomic Bulletin & Review*, *15*(4), 713–731. <https://doi.org/10.3758/PBR.15.4.713>

- Thierauf, A., Kempf, J., Eschbach, J., Auwärter, V., Weinmann, W., & Gnann, H. (2013). A case of a distinct difference between the measured blood ethanol concentration and the concentration estimated by Widmark's equation. *Medicine, Science and the Law*, *53*(2), 96–99. <https://doi.org/10.1258/msl.2012.012038>
- Tulving, E. (1972). Episodic and semantic memory. In E. Tulving & W. Donaldson (Eds.), *Organization of memory* (pp. 381–403). Academic.
- Tulving, E. (2002). Episodic memory: From mind to brain. *Annual Review of Psychology*, *53*, 1–25. <https://doi.org/10.1146/annurev.psych.53.100901.135114>
- Tulving, E., & Psotka, J. (1971). Retroactive inhibition in free recall: Inaccessibility of information available in the memory store. *Journal of Experimental Psychology*, *87*(1), 1–8. <https://doi.org/10.1037/h0030185>
- Tyson, P. D., & Schirmuly, M. (1994). Memory enhancement after drinking ethanol: Consolidation, interference, or response bias? *Physiology & Behavior*, *56*(5), 933–937. [https://doi.org/10.1016/0031-9384\(94\)90326-3](https://doi.org/10.1016/0031-9384(94)90326-3)
- Underwood, B. J. (1957). Interference and forgetting. *Psychological Review*, *64*(1), 49–60. <https://doi.org/10.1037/h0044616>
- Varma, S., Takashima, A., Krewinkel, S., van Kooten, M., Fu, L., Medendorp, W. P., Kessels, R. P. C., & Daselaar, S. M. (2017). Non-interfering effects of active post-encoding tasks on episodic memory consolidation in humans. *Frontiers in Behavioral Neuroscience*, *11*, Article 54. <https://doi.org/10.3389/fnbeh.2017.00054>
- Wamsley, E. J. (2019). Memory consolidation during waking rest. *Trends in Cognitive Sciences*, *23*(3), 171–173. <https://doi.org/10.1016/j.tics.2018.12.007>
- Weafer, J., Gallo, D. A., & De Wit, H. (2016a). Acute effects of alcohol on encoding and consolidation of memory for emotional stimuli. *Journal of Studies on Alcohol and Drugs*, *77*(1), 86–94. <https://doi.org/10.15288/jsad.2016.77.86>
- Weafer, J., Gallo, D. A., & De Wit, H. (2016b). Effect of alcohol on encoding and consolidation of memory for alcohol-related images. *Alcoholism: Clinical and Experimental Research*, *40*(7), 1540–1547. <https://doi.org/10.1111/acer.13103>
- White, K. G. (2001). Forgetting functions. *Animal Learning & Behavior*, *29*(3), 193–207. <https://doi.org/10.3758/BF03192887>
- Wickelgren, W. A. (1977). *Learning and memory*. Prentice-Hall.
- Wixted, J. T. (2004). The psychology and neuroscience of forgetting. *Annual Review of Psychology*, *55*, 235–269. <https://doi.org/10.1146/annurev.psych.55.090902.141555>
- Wixted, J. T., & Ebbesen, E. B. (1991). On the form of forgetting. *Psychological Science*, *2*(6), 409–415. <https://doi.org/10.1111/j.1467-9280.1991.tb00175.x>

- Wixted, J. T., & Ebbesen, E. B. (1997). Genuine power curves in forgetting: A quantitative analysis of individual subject forgetting functions. *Memory & Cognition*, *25*(5), 731–739. <https://doi.org/10.3758/BF03211316>

A Acknowledgements

This thesis is the result of a four-year journey that started in October 2020. As much as I enjoyed this time, it was also a period of challenges that I would not have endured without the guidance and support from so many wonderful people.

First of all, I would like to thank Edgar for supervising my dissertation project. You sparked my academic passion for memory research in the very first semester of my Bachelor studies in 2014. From becoming a student assistant at your chair in 2015, writing my Bachelor and Master theses under your supervision in 2018 and 2020, to now completing my dissertation, you have always been an incredible mentor. I am very thankful for your support and open-mindedness that allowed me to pursue my research interests.

I would like to thank Bea and Mandy for serving as additional supervisors. Our discussions were always very helpful for advancing my dissertation project. I also thank Arndt for our fruitful discussions in our colloquium and for agreeing to serve as thesis reviewer together with Bea.

I would like to thank all my dear colleagues and friends from the University of Mannheim and the SMiP group. I enjoyed our joint conference visits, retreats, workshops, and informal meetings very much and will always remember the great times we shared as doctoral students.

Finally, I would like to thank my family for their great support. I am particularly grateful to my wife Christina. Thank you for being my best friend.

B Statement of Originality

1. I hereby declare that the presented doctoral dissertation with the title *The Role of Consolidation in Interference-Based Forgetting: A Critical Re-Evaluation of Behavioral Evidence* is my own work.
2. I did not seek unauthorized assistance of a third party and I have employed no other sources or means except the ones listed. I clearly marked any quotations derived from the works of others.
3. I did not yet present this doctoral dissertation or parts of it at any other higher education institution in Germany or abroad.
4. I hereby confirm the accuracy of the declaration above.
5. I am aware of the significance of this declaration and the legal consequences in case of untrue or incomplete statements.

I affirm in lieu of oath that the statements above are to the best of my knowledge true and complete.

Signature:

Date:

C Co-Authors' Statements

Co-Author: Edgar Erdfelder

With this statement, I confirm that the following manuscripts included in the present thesis were primarily conceived and written by Julian Quevedo Pütter:

Quevedo Pütter, J., & Erdfelder, E. (2022). Alcohol-induced retrograde facilitation? Mixed evidence in a preregistered replication and encoding-maintenance-retrieval analysis. *Experimental Psychology*, 69(6), 335-350. <https://doi.org/10.1027/1618-3169/a000569>

Quevedo Pütter, J., Dahler, S., & Erdfelder, E. (2024). *Opportunistic consolidation or temporal distinctiveness? Retrieval, not storage, drives the temporal gradient of retroactive interference in episodic memory*. Manuscript submitted for publication.

Quevedo Pütter, J., & Erdfelder, E. (2024). *Waking rest during retention facilitates memory consolidation, but so does social media use: A storage-retrieval analysis*. Manuscript submitted for publication.

I sign this statement to the effect that Julian Quevedo Pütter is credited as the primary source of the ideas and the main author of the three above-listed manuscripts. He designed and programmed all experiments, and conducted the data collection for the experiment in Quevedo Pütter and Erdfelder (2022), Experiments 1 and 2 in Quevedo Pütter, Dahler, and Erdfelder (2024), and Experiment 2 in Quevedo Pütter and Erdfelder (2024). He analyzed all the data from all experiments, and wrote the first drafts and revisions of all three manuscripts. I contributed to the development and refinement of the research questions, study designs, and analyses. I also revised all three manuscripts.

Prof. Dr. Edgar Erdfelder
Mannheim, July 2024

Co-Author: Selina Dahler

With this statement, I confirm that the following manuscript included in the present thesis was primarily conceived and written by Julian Quevedo Pütter:

Quevedo Pütter, J., Dahler, S., & Erdfelder, E. (2024). *Opportunistic consolidation or temporal distinctiveness? Retrieval, not storage, drives the temporal gradient of retroactive interference in episodic memory*. Manuscript submitted for publication.

I sign this statement to the effect that Julian Quevedo Pütter is credited as the primary source of the ideas and the main author of the above-listed manuscript. He designed and programmed all three experiments, conducted the data collection for Experiments 1 and 2, contributed to the data collection of Experiment 3, analyzed all data from all three experiments, and wrote the first draft and revisions of the manuscript. I contributed to the design of Experiment 3, conducted most of the data collection for Experiment 3, and contributed to revising the manuscript.

Selina Dahler
Mannheim, July 2024

D Copies of Manuscripts



Alcohol-Induced Retrograde Facilitation?

Mixed Evidence in a Preregistered Replication and Encoding-Maintenance-Retrieval Analysis

J. Quevedo Pütter  and E. Erdfelder 

Department of Psychology, School of Social Sciences, University of Mannheim, Mannheim, Germany

Abstract: Somewhat counterintuitively, alcohol consumption following learning of new information has been shown to enhance performance on a delayed subsequent memory test. This phenomenon has become known as the retrograde facilitation effect (Parker et al., 1981). Although conceptually replicated repeatedly, serious methodological problems are associated with most previous demonstrations of retrograde facilitation. Moreover, two potential explanations have been proposed, the interference and the consolidation hypothesis. So far, empirical evidence for and against both hypotheses is inconclusive (Wixted, 2004). To scrutinize the existence of the effect, we conducted a preregistered replication that avoided common methodological pitfalls. In addition, we used Küpper-Tetzel and Erdfelder's (2012) multinomial processing tree (MPT) model to disentangle encoding, maintenance, and retrieval contributions to memory performance. With a total sample size of $N = 93$, we found no evidence for retrograde facilitation in overall cued or free recall of previously presented word pairs. In line with this, MPT analyses also showed no reliable difference in maintenance probabilities. However, MPT analyses revealed a robust alcohol advantage in retrieval. We conclude that alcohol-induced retrograde facilitation might exist and be driven by an underlying retrieval benefit. Future research is needed to investigate potential moderators and mediators of the effect explicitly.

Keywords: retrograde facilitation, alcohol, interference hypothesis, consolidation hypothesis, conceptual replication, multinomial processing tree (MPT) modeling



Alcohol is one of the most widely used and abused substances worldwide, with estimated prevalences of current drinking and heavy episodic drinking as high as 47% and 20%, respectively (Manthey et al., 2019). To better understand the consequences of drinking alcohol on everyday functioning, considerable effort has been devoted toward investigating cognitive alterations following acute alcohol intake (Mintzer, 2007). One particularly surprising observation originating from this research was that alcohol consumption following learning of new information apparently enhances performance on a delayed subsequent memory test. Parker et al. (1980) were the first to empirically demonstrate this so-called retrograde facilitation effect in a laboratory experiment using human subjects. While the effect seems to replicate across a variety of experimental paradigms, there are important methodological weaknesses associated with most of the studies used so far. Moreover, the question of which mechanisms underlie retrograde

facilitation has rarely been addressed empirically, and the corresponding research has not resulted in conclusive results so far (Wixted, 2004). Both issues are intertwined and of great importance for both theoretical and practical reasons. Therefore, in this registered report, we proposed a replication study that addresses them jointly. Specifically, we suggested a study design that avoids common methodological pitfalls by conceptually replicating the original procedure described by Parker et al. (1981) and, in addition, reanalyzing the replication data using a multinomial processing tree (MPT) model proposed by Küpper-Tetzel and Erdfelder (2012) to disentangle encoding, maintenance, and retrieval contributions to memory performance (for a review of MPT models, see Erdfelder et al., 2009; for an MPT tutorial, see Schmidt et al., 2023). Thus, we addressed two main goals in the present study: (a) rigorous conceptual replication of the retrograde facilitation effect to ensure that it is not an artifact of confounding variables and (b) using an established MPT measurement model to empirically assess the hypotheses proposed to explain retrograde facilitation. The preregistered study protocol is publicly available on the Open Science Framework (OSF; <https://doi.org/10.17605/OSF.IO/DK8FJ>; Quevedo Pütter et al., 2020).

Prior Research on Retrograde Facilitation

In 1981, Parker and colleagues demonstrated retrograde facilitation in a highly rigorous manner. In their experiment, alcohol versus placebo administration after learning was manipulated as a within-subjects factor. The retention interval lasted 7 hrs in each case, during which participants remained in a controlled laboratory environment, abstaining from any cognitively challenging tasks and further alcohol consumption. Participants showed significantly better performances in a recognition task after consuming either 0.5 or 1.0 ml alcohol per kg body weight (ml/kg) than after consuming 0.25 ml/kg or no alcohol at all. To illustrate, assuming a body weight of 80 kg, an alcohol weight of 0.8 g/ml, and an alcohol content of 5% for beer, the dose of 0.25 ml/kg would correspond to 20 ml (16 g) of pure alcohol or 0.4 l of beer, the dose of 0.5 ml/kg to 40 ml (32 g) of pure alcohol or 0.8 l of beer, and the dose of 1.0 ml/kg to 80 ml (64 g) of pure alcohol or 1.6 l of beer. Note that these quantities vary as a function of the body weights of individual participants and are thus examples only. Also note that beer is only used as a commonly known reference standard here. Parker et al. (1981) actually used a mixture of pure alcohol and a masking solution in their experiment.

In terms of internal validity, the Parker et al. (1981) study can be considered a convincing test of retrograde facilitation by alcohol. Most importantly, the retention interval of 7 h within a single day is (a) long enough to ensure that participants are completely sober when memory is tested and (b) short enough so that the interval does not cover a night of sleep. Unfortunately, other researchers have tried to avoid long retention intervals in controlled environments by either employing considerably shorter retention intervals (Knowles & Duka, 2004; R. E. Mann et al., 1984; Parker et al., 1980, study 1; Tyson & Schirmuly, 1994) or by dismissing participants after the learning phase and extending the retention interval up to 48 h (Bruce & Pihl, 1997; Carlyle et al., 2017; Lamberty et al., 1990; Mueller et al., 1983; Parker et al., 1980, study 2; Weafer et al., 2016a, 2016b). As detailed below, both approaches have considerable drawbacks compared to the original procedure used by Parker et al. (1981).

Short retention intervals involve the risk that participants are still intoxicated when memory is tested. To illustrate, R. E. Mann et al. (1984) conducted two separate studies with retention intervals of approximately 80 min and 140 min. They reported that participants were intoxicated to a considerable degree when memory was tested and rightfully concluded that memory tests conducted in a sober state for both the placebo and alcohol groups could have led to different results. One obvious criticism is that memory performance in the alcohol condition is hampered under such circumstances because

of state-dependent learning (Goodwin et al., 1969). Accordingly, memory performance of participants in the alcohol condition probably suffered from discrepant internal states during learning (sober state) and testing (intoxicated state). Nonetheless, statistically significant retrograde facilitation effects emerged in both studies of R. E. Mann et al. (1984). In principle, this finding is in line with a strong and robust retrograde facilitation effect. However, as participants in the alcohol condition were sober during learning and intoxicated during testing, the true size of the effect is probably underestimated, making it difficult to infer anything about its practical relevance. Even more importantly, as detailed below, one potential explanation of retrograde facilitation by alcohol – the interference hypothesis – can only be tested in a strict manner if accessibility of information in memory is not hampered by unrelated factors such as the state-dependency of memory.

An alternative approach to avoid long retention intervals in supervised controlled environments involves use of retention intervals extending to the next day. Hence, learning and memory testing take place on subsequent days, and participants can be sent home after the first session. This procedure has been employed by various researchers, with retention intervals ranging between 16 hrs (Carlyle et al., 2017) and 48 hrs (Weafer et al., 2016a, 2016b), always encompassing at least 1 night of sleep. While this approach avoids the problem of ongoing alcohol intoxication during recall, it comes with two other, possibly even more severe methodological challenges.

First, sleep research has shown that a presleep dose of alcohol typically increases slow-wave sleep (SWS) in the first half of the night for both healthy adults (Ebrahim et al., 2013) and adolescents (Chan et al., 2013). This poses serious problems for retrograde facilitation research because sleep, and especially SWS, has been argued to play a predominant role in the consolidation of declarative memories (Rasch & Born, 2013). Accordingly, it is possible that the apparent effects of retrograde facilitation by alcohol in these studies are actually due to memory consolidation during sleep. Alternatively, it could also be argued that effects of alcohol on sleep architecture may lead to a decline in memory performance. For example, an increase in sleep disruption following a presleep dose of alcohol (Ebrahim et al., 2013) could deteriorate participants' test performance on the following day due to sleep deprivation and increased daytime sleepiness during testing. However, this objection was taken into account in two studies by Weafer et al. (2016a, 2016b) who used a retention interval of 48 h to rule out that participants perform memory tests in a sleep-deprived or hungover state.

Second, acute alcohol intake has been shown to interact with sleep deprivation to increase daytime sleepiness

(Roehrs & Roth, 2001). Such an effect may be linked to the first problem, as daytime napping during the retention interval has been found to increase declarative memory performance (Tucker et al., 2006).

Mechanisms Proposed to Underlie Retrograde Facilitation by Alcohol

Two explanations for retrograde facilitation by alcohol have been proposed, the interference hypothesis and the consolidation hypothesis (Mueller et al., 1983; Tyson & Schirmuly, 1994). According to the interference hypothesis, retrograde facilitation is due to new incoming information being encoded and stored in memory less efficiently during intoxication, resulting in anterograde amnesia or at least mild forms thereof. As a consequence, memory representations originating from the previous learning phase are protected from retroactive interference. It has been shown convincingly that deterioration of memory following similarity-based retroactive interference is not due to distortion or loss of memory representations, but rather caused by impaired accessibility of information available in memory (e.g., Lohnas et al., 2015; Tulving & Psotka, 1971). By implication, the interference hypothesis predicts that retrograde facilitation results from enhanced accessibility of learned information in the testing phase, not from improved maintenance across the retention interval. In contrast, proponents of the consolidation hypothesis argue that alcohol actively boosts postlearning processing, resulting in more stable and durable memory representations (Parker et al., 1981), for example, by acting on brain regions involved in learning and memory such as the hippocampus (White, 1996) or by providing ideal (i.e., interference-free) conditions for undisrupted memory consolidation (Mednick et al., 2011; Wixted, 2004, 2010). Hence, retrograde facilitation emerges as a consequence of enhanced maintenance across the retention interval. In sum, while the interference hypothesis posits that alcohol-induced retrograde facilitation effects are due to better retrieval as a consequence of reduced retroactive interference, the consolidation hypothesis identifies better maintenance in memory as the main cause.

Both hypotheses have received some empirical support so far. Specifically, the interference hypothesis is in line with the findings by Mueller et al. (1983) and Tyson and Schirmuly (1994), respectively, showing that (a) retrograde facilitation is not time-dependent so that its strength does not depend on whether alcohol administration directly follows the learning phase or occurs later (although it has been shown that retroactive interference is also time-dependent to a certain degree; for a review, see Wixted, 2004) and (b) that retrograde facilitation is more

pronounced in memory tasks with few or no retrieval cues (e.g., free recall tasks) compared to tasks providing strong retrieval cues (e.g., recognition tasks), that is, effect sizes between alcohol and placebo conditions are more pronounced in free recall tasks than in recognition tasks. In contrast, Parker et al. (1981) favor a consolidation explanation because they observed a retrograde facilitation effect although a relatively low dose of alcohol was used in one condition (0.5 ml/kg) that is unlikely to decrease retroactive interference significantly. Perhaps more importantly, cognitive activity during the retention interval was reduced to a minimum in their study, making protection against interfering information obsolete.

Measurement of Memory Maintenance and Retrieval Contributions

Based on the evidence summarized in the previous section, Wixted (2004) concluded that attempts to empirically discriminate between the interference and the consolidation hypothesis have been inconclusive so far. He supposed that identification of the mediating physiological mechanisms might be a precondition for discriminating between the two theoretical accounts successfully. Following a different rationale, we propose an appropriate MPT model (see Erdfelder et al., 2009, for a review) to disentangle maintenance and retrieval contributions to the retrograde facilitation effect on a functional rather than a physiological level.

One particularly promising MPT model for our purposes is the encoding-maintenance-retrieval (EMR) model (Küpper-Tetzel & Erdfelder, 2012), an extension of the previously proposed storage-retrieval model by Rouder and Batchelder (1998). Küpper-Tetzel and Erdfelder's EMR model measures encoding, maintenance, and retrieval contributions to overall memory performance. It requires a study design that involves study of paired associates, immediately followed by a cued recall task. After a retention interval, additional free and final cued recall tests are administered (free-then-cued-recall paradigm). As detailed below, these methodological requirements are easily combined with the typical retrograde facilitation research design, enabling us to decide empirically whether retrograde facilitation by alcohol – if it exists – is driven by (a) improved maintenance across the retention interval, (b) improved retrieval in delayed subsequent free recall, or (c) both. Crucially, as shown above, the interference hypothesis and the consolidation hypothesis clearly differ with respect to which of these processes are expected to underlie retrograde facilitation. Also, because these hypotheses are not mutually exclusive, it is conceivable that both are correct, as reflected in improvements of both maintenance and retrieval after alcohol consumption.

Table 1. Twelve observable event categories E_1 to E_{12} for a study design employing an initial cued recall immediately following the learning phase and a free-then-cued-recall in the recall phase

Initial cued recall	Final cued recall	Final free recall		
		Both words	Exactly one word	Neither word
Correct	Correct	E_1	E_2	E_3
	Incorrect	E_4	E_5	E_6
Incorrect	Correct	E_7	E_8	E_9
	Incorrect	E_{10}	E_{11}	E_{12}

Note. Twelve possible memory test outcomes in the 3-level free-then-cued-recall paradigm proposed by Küpper-Tetzel and Erdfelder (2012).

Because the EMR model relies on a study design encompassing one initial cued recall test (2 possible outcomes per studied word pair: correct vs. incorrect), one later free recall test (3 possible outcomes: 0, 1, or 2 words of a pair recalled), and one final cued recall test (2 possible outcomes: correct vs. incorrect), a total of $2 \times 3 \times 2 = 12$ observable outcome patterns E_1 to E_{12} can occur for each word pair, depending on participants' performance in the three memory tests (cf. Table 1). The model comprises seven latent parameters: the probability of successful encoding of an association (e), the probabilities of maintenance of stored associations across the retention interval given successful versus unsuccessful initial cued recall (m_s and m_u , respectively), the probabilities of successful retrieval in free (r_f) and cued recall (r_c), and finally, the probabilities of single word retrieval in free recall given successful versus unsuccessful associative encoding (s and u , respectively). In total, the model includes 32 possible sequences of successful versus unsuccessful encoding, maintenance, and retrieval steps (so-called branches), each terminating in one of the 12 event categories E_1 to E_{12} . The probability of a branch is just the product of parameters along that branch, and the probability of a category equals the sum of the branch probabilities corresponding to this category. Based on these rules, model equations are obtained that represent the probabilities of the observable categories E_1 to E_{12} as functions of the seven model parameters. Given these equations and a set of event frequencies, model parameters can be estimated using either standard maximum likelihood techniques (Batchelder & Riefer, 1999; Hu & Batchelder, 1994; Moshagen, 2010) or Bayesian estimation techniques for hierarchical model versions that account for individual differences in model parameters (Heck et al., 2018). Küpper-Tetzel and Erdfelder (2012) have previously tested and validated the model successfully. Both an illustration of the model and the 12 model equations are available in the Electronic Supplemental Material (ESM) 1.

For our purposes, maintenance (m) and retrieval parameters (r_f) are key parameters to test the consolidation and the interference hypothesis, respectively. As detailed above, the interference hypothesis predicts that retrograde

facilitation results from enhanced accessibility of learned information in the testing phase. In the EMR model, this process is represented by the retrieval parameter r_f , the probability of successful retrieval of stored associations in free recall. In contrast, parameter r_c represents the probability of successful retrieval of the second word when the first word of an association stored in memory is provided as a cue. Given that the word pair is stored in memory, this cued recall probability should generally be very close to 1 in all conditions, as was previously found by Küpper-Tetzel and Erdfelder (2012). Parameter r_c is thus irrelevant for testing our hypotheses.

The consolidation hypothesis predicts that retrograde facilitation emerges as a consequence of enhanced maintenance across the retention interval. This process is represented by the maintenance parameters m_s and m_u . Küpper-Tetzel and Erdfelder (2012) observed that m_s and m_u can be equated to a single maintenance parameter m . Accordingly, the consolidation hypothesis predicts that m is larger in the alcohol condition compared to the placebo condition.

Parameter e should not be affected by alcohol administration after learning as it represents successful encoding of an association, a process that takes place before the experimental manipulation. Finally, parameters s and u represent probabilities of successful single word maintenance and retrieval for associations stored versus not stored in memory, respectively. They are thus not informative with respect to the question whether alcohol specifically enhances associative maintenance or retrieval.

It is important that the two parameters of prime interest, m and r_f , are not influenced by factors other than the administration of alcohol versus a placebo. It thus needs to be assured that (a) participants are completely sober when memory is tested after the retention interval to avoid contaminations of r_f with state-dependent learning decrements and (b) the retention interval does not include a night of sleep or daytime naps to preclude contaminations of successful maintenance (m) and retrieval (r_f) with sleep-related differences between conditions. For this purpose, we largely followed the original study by Parker et al. (1981) and employed a retention interval of 7 h that

ensured a sober state at both study and test. Additionally, the procedure was conducted in the controlled environment of a laboratory to preclude sleep, further consumption of alcohol, and other confounding variables. Note, however, that several methodological differences between our study and the original Parker et al. (1981) study remain. First, we did not restrict cognitive activity during the retention interval in any way (apart from enforcing compliance with the instructions). According to the interference hypothesis, the more cognitive activity during retention, the more accessibility benefits should emerge in the alcohol condition. Thus, a fair test of the interference hypothesis requires considerable cognitive activity in both conditions. Second, we used cued and free recall tasks for paired associates rather than assessing recognition memory for pictures. This adjustment is required by the EMR model. The same holds for the initial cued recall test immediately following learning, a procedure that deviates from Parker et al. (1981) but is mandatory for the EMR model-based analysis. Third, alcohol consumption was manipulated as a between-subjects factor in our study. We feel that this prevents alcohol-related expectancy effects more effectively than a within-subjects manipulation because participants lack a direct reference standard for the alcohol content. Fourth, we did not restrict our study to male participants but investigated all genders. We took into account the difference in alcohol tolerance between genders by administering different doses to women (0.51 g/kg) and men (0.59 g/kg).

Hypotheses

Because the retrograde facilitation effect emerged in all prior studies we were aware of, even in those in which its strength was probably hampered by state-dependent learning influences, we assumed that the effect truly exists and thus hypothesized that forgetting in cued recall across the retention interval (i.e., the difference in correct responses between the initial and the final cued recall) would be reduced for participants in the alcohol condition compared to the placebo condition (Hypothesis 1) and, in addition, that participants in the alcohol condition would retrieve more word associations in final free recall (Hypothesis 2).

Moreover, in line with the interference hypothesis, we hypothesized that the EMR model probability of retrieving word associations in free recall (parameter r_f) would be significantly higher for participants in the alcohol condition (Hypothesis 3). Based on the consolidation hypothesis, we additionally hypothesized that maintenance of stored word associations across the retention interval (parameter m) would be significantly higher for participants in the

alcohol condition (Hypothesis 4). No specific predictions applied to other model parameters.

Method

All methodological details of the present study were planned and evaluated by the authors in accordance with the ethical principles outlined in the Declaration of Helsinki (2013). The research protocol was approved by the ethics committee of the University of Mannheim. The materials necessary to replicate this study are available on the OSF (<https://doi.org/10.17605/OSF.IO/8E9PW>).

Design

The present study was conducted in a double-blind, randomized, placebo-controlled manner. Alcohol administration was a between-subjects factor. Thus, one group of participants consumed an individually determined dose of alcohol after learning while another group of participants received a perceptually indistinguishable placebo beverage. Participants were randomly assigned to one of the two conditions.

Participants

Sample Size

We defined a medium effect size of $d = 0.50$ as the minimum effect of interest in the present study. With $d = 0.50$, $\alpha = .05$, and a desired statistical power of $1 - \beta = .80$, a conventional Neyman-Pearson power analysis using the software G*Power (Faul et al., 2007) results in a required sample size of $N = 102$ for a one-tailed (i.e., directed) 2-groups t test (51 per experimental condition). To maximize efficiency, a sequential t test (e.g., Schnuerch & Erdfelder, 2020) was used to test the differences in cued and free recall memory performance between conditions, an approach that has been shown to reduce the required sample size to about 60% of the corresponding Neyman-Pearson sample size on average (see below). If both sequential tests would terminate with $n < 30$ in either condition, additional participants would be sampled subsequently until the threshold $n = 30$ would have been reached in each condition to enable meaningful MPT analyses of the data.

Recruitment

Potential participants were recruited via the online platform for study participation of the University of Mannheim and by advertisement within the university campus. Participants received study credit. The top three

performances in the immediate cued recall and the final free-then-cued-recall were rewarded with 20 € each. Study information advised subjects to participate in groups because they would watch movies, eat pizza, and drink alcohol as part of a study taking 8.5 hrs in total and allegedly investigating processing of movies under the influence of alcohol.

Eligibility

Interested individuals were required to report at least one heavy drinking episode (five alcoholic beverages for men and four alcoholic beverages for women at one occasion; Wechsler et al., 1995) within the last month to ensure that they had sufficient experience to tolerate the to-be-consumed alcohol dose. Additional inclusion criteria included age between 18 and 29 years, body mass index (BMI) between 18.5 and 29.9, no current or past diagnosis of a substance use disorder or any other addictive disorder, no current diagnosis of any other psychiatric disorder, no physical disorder that precludes the consumption of alcohol, no medical advice to avoid the consumption of alcohol, no current intake of any medication other than birth control, no pregnancy or possibility of pregnancy, and no lactation. Age and BMI restrictions were included to allow precise estimation of blood alcohol concentrations (BACs) by the Widmark formula (Thierauf et al., 2013). All criteria were assessed by self-report. Social drinking and no substance use disorder were further confirmed through the Alcohol Use Disorders Identification Test (AUDIT; see below).

Materials

Alcohol and Placebo

For participants assigned to the alcohol condition, the alcohol dose was 0.51 g/kg for women and 0.59 g/kg for men.¹ Based on the Widmark formula (Widmark, 1932, as cited in Thierauf et al., 2013) and following the refined procedure outlined by Thierauf et al. (2013), these doses could be expected to result in BACs of around 0.60‰ (see ESM 2), an intoxication level high enough to cause significant cognitive impairment. Assuming an elimination rate of around 0.15‰ per hour (Thierauf et al., 2013), it could be expected that alcohol would be eliminated from the body within the retention interval of 7 h. Moreover, the doses lie within the range of dosages that have successfully

been employed in previous research on retrograde facilitation, ranging from 0.40 g/kg (Parker et al. 1981) to 0.80 g/kg (e.g., Weafer et al., 2016a). The difference between the two doses for women and men accounts for the fact that women typically exhibit higher BACs than men for a given amount of consumed alcohol (Fillmore, 2001).

In the alcohol condition, a mixture of vodka, tonic water, and Tabasco sauce was served as described by Knowles and Duka (2004). The beverage in the alcohol condition was made up of one part vodka (40% alcohol content) and three parts tonic water such that the total amount of alcohol reached the individually predetermined dose. For administration, each individual beverage was split into 10 portions, meaning that every participant was asked to consume their respective total beverage in 10 smaller portions. To mask the taste and burn of vodka, two drops of Tabasco sauce were added to each portion. In the placebo condition, vodka was replaced with additional tonic water. In both conditions, glass rims were swabbed with vodka to further intensify the sensual impression of alcohol in the placebo condition (Lamberty et al., 1990). Hence, participants in the placebo condition actually did consume some alcohol, but a very low and negligible dose. Due to the inclusion criterion of a BMI between 18.5 and 29.9, the total amount of individual beverages in the alcohol condition could range between 252 ml (67 ml vodka + 185 ml tonic water, split into 10 portions of 25.2 ml each) in case of female gender and body weight of 42 kg, and 832 ml (221 ml vodka + 611 ml tonic water, split into 10 portions of 83.2 ml each) in case of male gender and body weight of 120 kg. Note that these values are theoretical extremes that were not encountered in our study. Actual beverages varied between 266 ml (71 ml vodka + 195 ml tonic water, split into 10 portions of 26.6 ml each) and 624 ml (166 ml vodka + 458 ml tonic water, split into 10 portions of 62.4 ml each).

Word Pairs

Forty German word pairs were taken from Hager and Hasselhorn (1994), the same as used by Küpper-Tetzl and Erdfelder (2012). These word pairs have the important characteristic of being only weakly associated to another. Hence, the probability of successful generation of the target word by mere guessing when presented with the cue word in cued recall tasks is minimized. A complete overview of the learning material is provided in ESM 3.

¹ Due to an initial error in calculating alcohol doses, the doses administered in this study were slightly lower than preregistered for the first $n_1 = 28$ participants ($N_{\text{total}} = 93$), i.e., approximately 0.44 g/kg for women and 0.51 g/kg for men. For the remaining $n_2 = 65$ participants, the preregistered doses were administered. Crucially, this difference in doses did not result in a relevant difference in peak BAC measurements in the alcohol condition, $M_1 = 0.42\text{‰}$ ($SD = 0.08$), $M_2 = 0.44\text{‰}$ ($SD = 0.08$), $t(44) = 0.60$, $p = .275$, Cohen's $d = 0.19$. Moreover, the lower-than-preregistered doses still lie above the medium dose of 0.40 g/kg in Parker et al. (1981), which was sufficient to yield a substantial retrograde facilitation effect.

AUDIT

A German translation of the Alcohol Use Disorders Identification Test (AUDIT; Saunders et al., 1993) was used to check the exclusion criterion of hazardous drinking habits and the inclusion criterion of social drinking. This scale is recommended in German clinical guidelines for the screening of alcohol use disorders (K. Mann et al., 2017). It consists of 10 items asking for the amount and frequency of drinking occasions and the occurrence of several negative consequences following drinking. Items include three to five alternatives that are rated with a score between 0 and 4. Thus, the maximum total score is 40. Total scores of 8 or more are treated as indicators of harmful alcohol use (Babor et al., 2001). Therefore, interested persons scoring above 7 were not allowed to participate. The inclusion criterion of social drinking was ensured by allowing participation only if individuals reached a total score of at least 2 on the first two items (i.e., one drinking occasion per month with at least three to four alcoholic drinks or two to four drinking occasions per month with at least one or two alcoholic drinks) and a score of at least 1 on the third item (i.e., heavy drinking episodes in the past).

Procedure

Individuals interested in participating in this study first accessed an online survey that provided them with a description of the study procedure. To prevent demand effects, participants were told that (a) all participants in this study would consume alcohol in a more or less high dose and that (b) the focus of this study lay on the perception and processing of movies under the influence of alcohol. Inclusion criteria were also checked at this occasion. Eligible participants were asked not to drink alcohol within 48 hrs prior to their participation in the experiment and to have lunch before participation began. The experiment took place in a classroom-like laboratory within the premises of the Department of Psychology at the University of Mannheim.

Participants were asked to arrive at 1:30 p.m. at the laboratory. With the exception of general information and instructions regarding the procedure, the first phase of the experiment was administered individually. First, participants were asked for inclusion criteria again and provided written consent. Additionally, breath alcohol concentration was measured to ensure sobriety of all participants. Then, participants were instructed for the learning phase. Forty word pairs were presented in randomized order for 5 s each on a computer screen, with participants being informed that memory for the materials would be tested afterward in a cued recall task. The announced immediate

cued recall was conducted following a short distractor task where participants were asked to sequentially assess the correctness of 15 equations as quickly as possible. In the immediate cued recall, participants were required to complete each cue word with the respective target word. Cue words were presented in randomized order on a computer screen with participants responding by filling in the respective target words at their own pace. After completion, participants were asked to drink 10 portions of a beverage (either the alcohol or the placebo beverage, depending on the condition they were assigned to) within 30 min (i.e., 3 min per portion). Both the experimenter and the participant were blind with respect to the experimental condition. Following the first portion, participants were asked to estimate the alcohol content of the beverage on a scale from “less than 1%” to “equal to or more than 20%” (note that the true value was about 10% for all participants in the alcohol condition) to check whether the placebo beverage successfully blinded participants for their experimental condition (Keane et al., 1980). After consuming the final portion, participants of both conditions were asked to rinse their mouth with water. Participants were instructed individually not to communicate with other participants about their experiences during learning and alcohol administration. This completed the first phase of this study that lasted approximately 1 h.

The retention interval of 7 h was spent by participants in a seminar room under permanent supervision. Because participants were encouraged to register in groups, they could freely interact and socialize with their colleagues during this time. As part of the cover story, movies were presented. As participation in the study covered the evening hours, pizza was provided for dinner. Additionally, nonalcoholic drinks were available all the time. Breath alcohol concentrations of all participants were checked 30, 60, and 90 min after the administration of alcohol or a placebo (to assess peak BACs) and immediately before memory testing began (to ensure sobriety of all participants). Participants were told to abstain from drinking alcohol and sleeping over the whole course of the retention interval. Compliance with the instruction not to communicate about the learning phase and the alcohol administration was monitored by the experimenter.

After the retention interval, participants performed a surprise free-then-cued-recall for the learned word associations. In the free recall task, participants were asked to recall as many of the word pairs as possible within 8 min. Importantly, participants were instructed to write down single words in cases when they did not remember both words of an association. Next, a final cued recall identical to the immediate cued recall task after the study phase was employed. Additionally, participants were asked to declare whether they followed the instruction not to communicate

about their experiences from the first phase of the experiment and whether they thought about or actively rehearsed any word pairs in a postexperimental questionnaire. Finally, participants were informed about the true theoretical background and design of the study and the exact dose of alcohol they consumed. Additionally, research assistants were present to answer further questions and concerns. The study ended by 10.00 p.m.

Data Analysis

All statistical tests were conducted with $\alpha = .05$. As a manipulation check, peak BACs in the alcohol condition were expected to be significantly larger than zero and significantly larger than BACs observed in the placebo condition. BACs in the placebo condition were expected never to rise to a level larger than zero. Blinding of participants for their condition would be regarded as successful if participants in the placebo condition estimated the alcohol content of their beverage as 1% or more, indicated by all responses other than “less than 1%”. Analyses were conducted with and without those participants in the placebo condition who estimated the alcohol content as less than 1% to test whether results were affected by BAC awareness. Similarly, analyses were conducted with and without those participants who communicated about Phase 1 of the experiment with other participants.

To maximize efficiency of data collection, the sampling process was conducted using a one-tailed group-sequential probability ratio (SPRT) t test. Thus, after each group of participants investigated as described above, the resulting data were analyzed to decide whether a decision for or against Hypothesis 1 could be made or further data were required. This sequential procedure has been shown to result in a considerable reduction of required sample sizes when compared to conventional Neyman-Pearson t tests. At the same time, error probabilities are controlled (Schnuerch & Erdfelder, 2020). In our application, standard parameters for the Type-1 and Type-2 error probabilities were used (i.e., $\alpha = .05$ and $\beta = .20$, respectively, assuming a medium effect size $d = .50$, cf. Cohen, 1988). Difference scores were calculated for every participant, subtracting the number of correct responses in the final cued recall from the number of correct responses in the immediate cued recall. Then, the difference between the two mean difference scores of experimental conditions were tested for statistical significance using the SPRT t test as specified above until a decision was made.

To test Hypothesis 2, the number of complete word pairs recalled was calculated for every participant to obtain mean performances for both experimental conditions. The difference of these two means was tested for statistical

significance with the same group-sequential one-tailed two-sample SPRT t test as in case of Hypothesis 1.

Hypotheses 3 and 4 referred to parameters r_f and m in the EMR model. Participants underperforming (<30% correct) or overperforming severely (>80% correct) in the immediate cued recall needed to be excluded from MPT analyses to ensure sufficient data points for event categories E_1 to E_6 and E_7 to E_{12} , respectively. A minimum sample size of $n = 30$ per condition was ensured to enable trustworthy parameter estimates for the EMR model in either of the conditions. Next, the frequencies of the 12 event categories E_1 to E_{12} were calculated individually for all remaining participants and aggregated within conditions. The resulting data could then be used to fit the EMR model (see above). All MPT analyses were conducted twice using TreeBUGS (Heck et al., 2018) for individual data and multiTree (Moshagen, 2010) for aggregated data. Following Küpper-Tetzel and Erdfelder (2012), we would (a) try to equate m_s and m_u so that a single parameter m represents the probability of successful maintenance and (b) equate r_c across experimental conditions. For the aggregated data, this leads to 2 (= number of conditions) \times 5 (= unrestricted model parameters) + 1 (= number of parameters equated for both conditions) = 11 model parameters to be estimated given 2 (= number of conditions) \times 11 (= number of free categories per condition) = 22 independent category frequencies. Accordingly, the number of degrees of freedom of the G^2 goodness-of-fit tests test is 22 – 11 = 11. We would accept the model fit if $G^2(11)$ would result in a p value greater than .05. If this criterion would not be met, the model would need to be replaced by a more general model version, for example, by allowing m , s , and/or u to differ between successful and unsuccessful immediate cued recall.

To test Hypothesis 3, an equality constraint would be imposed on the r_f parameters of both experimental conditions. If the resulting decrease in model fit $\Delta G^2(1)$ would become significant at $\alpha = .05$ (one-tailed) and parameter estimates would be in the direction predicted by the interference hypothesis, the latter hypothesis would be confirmed. For the test of Hypothesis 4, analogous procedures would be used: An equality constraint would be imposed on the m parameters of both experimental conditions, and the hypothesis would be confirmed if the resulting decrease in model fit $\Delta G^2(1)$ would become significant at $\alpha = .05$ (one-tailed) and parameter estimates would be in the predicted direction.

Results

The full data set used for the analyses and the corresponding codebook are provided on the OSF (<https://doi.org/10.6084/m9.figshare.13111111>).

org/10.17605/OSF.IO/2K4J7). Additionally, the R code to obtain the results reported below is provided in ESM 4.

Sample

A total of $N = 93$ participants (divided into 20 sequential groups, 2–10 participants per group, 46 participants in the alcohol condition) took part in the experiment. Although a decision for both SPRT t tests was already reached earlier (see below), it was not before this point of the data collection process that the preregistered minimum of $n = 30$ participants per condition was achieved for our encoding-maintenance-retrieval MPT analyses. In the full sample, 62 participants were female, and 31 participants were male. The mean age was 20.67 years ($SD = 2.18$). A majority of 91 participants reported to be university students, of which 72 participants were enrolled in a psychology program. The mean AUDIT score was 5.39 ($SD = 1.23$).

Control Variables and Manipulation Checks

BAC measurements at the beginning of the experiment confirmed that all participants arrived at the laboratory completely sober. Six participants reported to have consumed alcohol within 48 h before the experiment. Most participants in the alcohol condition (41 of 46 participants) reached their measured peak BAC 30 min after the end of the alcohol administration, four participants reached it after 60 min, and one participant after 90 min. The mean peak BAC in the alcohol condition was 0.43‰ ($SD = 0.08$, Range = 0.22–0.62). At all three measurement occasions after the alcohol administration, mean BACs in the alcohol condition were significantly larger than zero: After 30 min, $M_{\text{BAC}} = 0.43\text{‰}$ ($SD = 0.08$, Range = 0.22–0.62), $t(45) = 34.74$, $p < .001$; after 60 min, $M_{\text{BAC}} = 0.37\text{‰}$ ($SD = 0.08$, Range = 0.18–0.57), $t(45) = 31.40$, $p < .001$; and after 90 min, $M_{\text{BAC}} = 0.30\text{‰}$ ($SD = 0.08$, Range = 0.12–0.46), $t(42) = 24.91$, $p < .001$.² Eight of 47 participants in the placebo condition estimated the alcohol content of their beverage to be below 1%. Actual BACs in the placebo condition never rose to a level larger than zero, except for one participant who had a BAC measure of 0.08‰ 90 min after the end of the placebo administration. All participants in both conditions were completely sober again at the last BAC measurement immediately before the final recall tests.

In the postexperimental questionnaire, 23 participants (13 from the alcohol condition, 10 from the placebo

condition) reported to have communicated about their beverages, and two participants (one from the alcohol condition, one from the placebo condition) reported to have communicated about both the beverages and the word pairs during the retention interval. Thirty-three participants (14 from the alcohol condition, 19 from the placebo condition) reported to have thought about the word pairs during the retention interval. Of these, three participants (two from the alcohol condition, one from the placebo condition) reported to have engaged in active rehearsal of the word pairs.

Design-Based Results

Hypotheses 1 and 2 referred to the difference in correct responses between immediate and final cued recall and the number of complete word pairs reproduced during free recall, respectively. Both hypotheses were tested using the SPRT t test (Schnuerch & Erdfelder, 2020) implemented in the sprtt R package (v0.1.0; Steinhilber et al., 2021). For both Hypotheses 1 and 2, the decision to accept the null hypothesis of no significant alcohol benefit was reached rather quickly after $N = 18$ subjects (divided into five sequential groups, $n_{\text{Alcohol}} = n_{\text{Placebo}} = 9$) had participated. The means and standard deviations from this subsample and the full sample ($N = 93$) for the immediate cued recall, the final cued recall, the cued recall difference (Hypothesis 1), the number of complete word pairs reproduced during free recall (Hypothesis 2), and the number of single words reproduced during free recall are reported in Table 2.

For Hypothesis 1, the likelihood ratio (LR) of the SPRT t test at $N = 18$ was $LR_{18} = 0.18$, thereby undercutting the lower SPRT threshold implied by our preregistered SPRT parameters, that is, $\beta/(1 - \alpha) = 0.20/0.95 = 0.21$. The observed LR indicates that the data at $N = 18$ were about $1/0.18 = 5.6$ times more likely under H_0 than under H_1 , thus enforcing acceptance of H_0 . The sample estimate of Cohen's d was -0.53 , that is, there was a medium-sized effect in the direction opposite to Hypothesis 1. For Hypothesis 2, we observed $LR_{18} = 0.11$. Thus, at $N = 18$, the data were about $1/0.11 = 9.1$ times more likely under H_0 than under H_1 , also enforcing acceptance of H_0 . In this case, there was a strong effect in the direction opposite to Hypothesis 2, Cohen's $d = -0.82$. In sum, no significant retrograde facilitation effect could be observed for either of our two dependent variables. Plots of the developments of the log-likelihood ratios for both SPRT t tests are provided in ESM 5.

² Due to technical problems during the experiment, BAC measurements are missing for three participants in the alcohol condition 90 min after the end of the alcohol administration and immediately before the final recall tests.

Table 2. Means and standard deviations (SD) of all dependent variables in the alcohol and placebo condition

Dependent variable	SPRT subsample (N = 18)		Full sample (N = 93)	
	Alcohol	Placebo	Alcohol	Placebo
Immediate cued recall	19.33 (7.94)	26.56 (4.64)	23.57 (8.89)	23.83 (7.50)
Final cued recall	16.78 (8.80)	24.89 (4.83)	21.85 (9.59)	21.34 (7.65)
Cued recall difference	2.56 (1.81)	1.67 (1.50)	1.72 (1.73)	2.49 (2.41)
Free recall: complete pairs	9.22 (4.60)	13.22 (5.09)	11.57 (5.44)	10.04 (4.66)
Free recall: single words	23.67 (8.89)	31.56 (9.89)	28.11 (10.68)	26.21 (9.29)

Note. Decisions in favor of H_0 for Hypotheses 1 (cued recall difference) and 2 (free recall: complete pairs) were reached by the sequential probability ratio t tests (SPRT t tests) after data from $N = 18$ participants had been collected. Due to the second stopping criterion of $n = 30$ in both conditions for the multinomial processing tree (MPT) analysis (after excluding participants over- or underperforming in the immediate cued recall), a total of $N = 93$ subjects participated in this study. *SDs* in parentheses.

Although irrelevant for the SPRT t test decisions in favor of H_0 , we additionally inspected the likelihood ratios of the full-sample SPRT t tests, LR_{93} , for explorative reasons. Moreover, because LR_{93} assumes a fixed effect size of $d = 0.50$ under H_1 – a debatable assumption – we additionally computed corresponding Bayes factors (BF_{10}) using the default Cauchy prior as implemented in the BayesFactor R package (v0.9.12-4.2; Morey & Rouder, 2018). As summarized in Table 2, the descriptive pattern regarding our two dependent variables of interest was reversed in the full sample of $N = 93$ as compared to the subsample of $N = 18$ required until termination of the SPRT. However, likelihood ratios in the full sample indicate no clear evidence in favor of H_1 in both cases; $LR_{93} = 3.84$, Cohen's $d = 0.37$ for Hypothesis 1 and $LR_{93} = 1.81$, Cohen's $d = 0.30$ for Hypothesis 2. In line with this, Bayesian t tests revealed $BF_{10} = 1.63$ for Hypothesis 1 and $BF_{10} = 1.00$ for Hypothesis 2. Thus, although Hypotheses 1 and 2 are descriptively in line with the data of the full sample, the evidence in favor of H_1 is negligible in both cases (Kass & Raftery, 1995).

To further scrutinize the robustness of our conclusions regarding Hypotheses 1 and 2, we excluded participants from both the subsample and the full sample who either (a) were assigned to the placebo condition and estimated the alcohol content of their beverage to be less than 1% or (b) reported to have communicated about Phase 1 of the experiment with other participants. The means and standard deviations for the two dependent variables of interest are reported in Table 3.

After the application of both exclusion criteria, $N = 12$ participants remained in the subsample, whereas $N = 62$ participants remained in the full sample. The descriptive patterns mirror those obtained without consideration of these exclusion criteria. Similarly, corresponding likelihood ratios and Bayes factors support the conclusion of no clear evidence in favor of H_1 : For Hypothesis 1, $LR_{12} = 0.39$, $LR_{62} = 2.54$, $BF_{10} = 1.15$; for Hypothesis 2, $LR_{12} = 0.26$, $LR_{62} = 1.53$, $BF_{10} = 0.81$.

Model-Based Results

Hypotheses 3 and 4 referred to the retrieval parameter r_f and the maintenance parameter m of the EMR model, respectively. For the aggregated data, we conducted all analyses in multiTree (Moshagen, 2010). The corresponding preregistered analyses are provided in ESM 6. The baseline EMR model as described in the Data Analysis section did not meet our model fit criterion with $\alpha = .05$, $G^2(11) = 22.00$, $p = .024$, AIC = 9,758.71, BIC = 9,824.18. Therefore, we defined a generalized model version with parameters m , u , and s allowed to vary within both conditions between successful and unsuccessful immediate cued recall (m_s and m_u , u_s and u_u , and s_s and s_u , respectively). This generalized model fit the data well, $G^2(5) = 5.97$, $p = .309$, AIC = 9,754.68, BIC = 9,855.85. A close inspection of this model revealed a significant difference between parameters u_s and u_u only within the placebo condition, $\Delta G^2(1) = 12.55$, $p < .001$, but not within the alcohol condition, $\Delta G^2(1) = 1.06$, $p = .303$. Thus, participants in the placebo condition had a significantly higher probability of single word retrieval in the free recall for unsuccessfully stored associations after successful immediate cued recall ($u_s = .21$) than after unsuccessful immediate cued recall ($u_u = .10$), in contrast to participants in the alcohol condition ($u_s = .16$, $u_u = .12$).

Accordingly, we defined a new baseline model with the following restrictions: In line with Küpper-Tetzl and Erdfelder (2012), parameters m_s and m_u as well as s_s and s_u were equated within both experimental conditions, and parameter r_c was equated across conditions. In contrast, parameter u was allowed to vary freely within both conditions between successful (u_s) and unsuccessful (u_u) immediate cued recall. This redefined model yielded a good fit to the data, $G^2(9) = 7.43$, $p = .592$, AIC = 9,748.14, BIC = 9,825.51. As this model version is more parsimonious than the generalized model version, we decided to use it as our baseline model for the hypothesis tests.

Table 3. Means and standard deviations (SD) of both dependent variables of interest in the alcohol and placebo conditions after exclusion criteria were applied

Dependent variable	SPRT subsample ($N = 12$)		Full sample ($N = 62$)	
	Alcohol	Placebo	Alcohol	Placebo
Cued recall difference	2.67 (2.07)	2.00 (1.26)	1.91 (1.86)	2.73 (2.57)
Free recall: complete pairs	9.67 (4.84)	13.33 (5.75)	11.47 (5.97)	9.80 (4.92)

Note. Participants from the crucial sequential probability ratio test (SPRT) subsample and the full sample were excluded if they estimated the alcohol content of their beverage to be less than 1% in the placebo condition and if they reported to have communicated about Phase 1 of the experiment with other participants. SDs in parentheses.

This baseline model was additionally fitted to the individual data in the form of a Bayesian hierarchical model. For this purpose, the model was estimated for both conditions separately, using the latent trait framework of Klauer (2010) as implemented in TreeBUGS (Heck et al., 2018). In this Bayesian framework, convergence of parameter estimates can be evaluated by means of the potential scale reduction factor R (Gelman & Rubin, 1992). Good convergence was obtained for all parameters in the alcohol condition, $\hat{R} \leq 1.007$, and also in the placebo condition, $\hat{R} \leq 1.003$. The model fit was evaluated using the goodness-of-fit statistics T_1 and T_2 (Klauer, 2010). Good model fit was obtained for either condition, as indicated by posterior predictive p values of $p_1 = .288$ ($p_2 = .300$) in the alcohol condition and $p_1 = .565$ ($p_2 = .420$) in the placebo condition. In Bayesian hierarchical MPT modeling, statistical reliability of parameter differences is typically assessed by checking (a) whether the 95% Bayesian credibility interval (BCI) of the posterior distribution of the difference estimate does not include zero (two-tailed) or (b) whether the Bayesian p value, that is, the proportion of the posterior distribution of the difference estimate below zero is smaller than the chosen significance level of $\alpha = .05$ (one-tailed).

After excluding all participants who underperformed (<30% correct) or overperformed severely (>80% correct) in the immediate cued recall, a subsample of $N = 71$ participants remained for the preregistered model-based analysis (31 in the alcohol condition, 40 in the placebo condition). The parameter estimates for both the aggregated and the individual data are presented in Table 4.

As expected, the probability of associative encoding (parameter e) did not differ significantly between conditions, neither for the aggregated data, $\Delta G^2(1) = 0.55$, $p = .460$, nor for the individual data, 95% BCI = $[-.09, .06]$. Additionally, as expected, parameter r_c was estimated to be very close to 1 in both modeling approaches.

Based on the interference hypothesis, we predicted a significantly higher probability of associative retrieval in free recall (parameter r_f) for participants in the alcohol condition (Hypothesis 3). Parameter estimates of both modeling approaches were in line with this prediction, and this descriptive effect did reach statistical significance (one-tailed) both for

the aggregated data, $z = \sqrt{\Delta G^2(1)} = 2.35$, $p = .009$, and for the individual data, Bayesian $p = .023$.

Based on the consolidation hypothesis, we hypothesized the probability of associative maintenance (parameter m) to be significantly higher for participants in the alcohol condition (Hypothesis 4). Again, parameter estimates of both modeling approaches were descriptively in line with our prediction. However, this descriptive effect was statistically significant (one-tailed) only for the aggregated data, $z = \sqrt{\Delta G^2(1)} = 2.33$, $p = .010$, but not for the individual data, Bayesian $p = .104$.

Overall, the results of our preregistered MPT analysis were in line with Hypothesis 3, but did not yield clear evidence in favor of Hypothesis 4. Additionally, the descriptive effect for parameter m was rather small, irrespective of the modeling approach. To further clarify the robustness of this conclusion, we conducted an exploratory MPT multiverse analysis. In this analysis, we not only included our preregistered exclusion criteria of immediate cued recall performance, alcohol content estimation in the placebo condition, and communication during the retention interval (see the “Data Analysis” section), but additionally considered the influence of failure not to drink alcohol within 48 h prior to the experiment and active word pair rehearsal during the retention interval. Thus, our multiverse analysis had a 2 (modeling approach: aggregated vs. individual data) \times 2 (model-related exclusion criterion present vs. absent: under- or overperforming in the immediate cued recall) \times 2 (at least one procedure-related exclusion criterion present vs. absent: drinking alcohol within 48 h prior to the experiment, estimating the alcohol content to be below 1% in the placebo condition, communicating about the alcohol vs. placebo administration and/or the learning phase, actively rehearsing word pairs during the retention interval) design.

The first two cells of this multiverse design contained our preregistered analyses reported above, where both modeling approaches were applied to the data of all participants who did not meet the model-related exclusion criterion, whereas procedure-related exclusion criteria were not considered. The results of the remaining cells are reported in ESM 7. For both the aggregated and the individual data, the model fit the data well in all remaining

Table 4. Parameter estimates from the main multinomial processing tree (MPT) analysis

Parameter	Alcohol condition				Placebo condition			
	Aggregated data		Individual data		Aggregated data		Individual data	
	<i>MLE</i>	95% CI	<i>M</i>	95% BCI	<i>MLE</i>	95% CI	<i>M</i>	95% BCI
<i>e</i>	.58	[.55, .61]	.59	[.53, .64]	.60	[.57, .62]	.60	[.55, .65]
<i>m</i>	.92	[.90, .95]	.93	[.90, .96]	.89	[.86, .91]	.90	[.86, .93]
<i>r_c</i>	.98	[.97, .98]	.98	[.97, .99]	.98	[.97, .98]	.98	[.97, .99]
<i>r_f</i>	.50	[.46, .54]	.51	[.45, .56]	.44	[.41, .47]	.43	[.39, .48]
<i>s</i>	.05	[.03, .07]	.05	[.03, .07]	.11	[.09, .13]	.11	[.09, .14]
<i>u_s</i>	.16	[.08, .24]	.18	[.07, .34]	.22	[.15, .28]	.20	[.12, .28]
<i>u_d</i>	.12	[.10, .14]	.11	[.08, .15]	.10	[.09, .12]	.11	[.09, .13]

Note. For the aggregated data, the encoding-maintenance-retrieval (EMR) model was fitted using the multiTree software (Moshagen, 2010). Maximum likelihood parameter estimates (*MLE*) are presented alongside the corresponding 95% confidence interval (CI). Parameter *r_c* was equated across conditions. For the individual data, the model was fitted for both conditions separately using the R package TreeBUGS (Heck et al., 2018). Posterior means (*M*) are presented alongside the corresponding 95% Bayesian credibility intervals (BCI).

cells, thereby allowing for an interpretation of the respective parameter estimates. Overall, our conclusions for both Hypotheses 3 and 4 were confirmed: Parameter *r_f* differed reliably between conditions in all cells for the aggregated data (all $p \leq .034$) and in all cells except one (Bayesian $p = .051$) for the individual data (both other Bayesian $p \leq .016$). In contrast, parameter *m* differed reliably in only one additional cell ($p = .010$) for the aggregated data (both other $p \geq .090$) and in no cell for the individual data (all Bayesian $p \geq .132$).

Discussion

In the present study, we attempted to replicate the counterintuitive phenomenon of alcohol-induced retrograde facilitation. Using a retention interval of 7 hrs, we followed the basic procedure introduced by Parker et al. (1981) in their original study. Thus, we avoided possible confounds of interpolated sleep and other memory-relevant activities and thereby put the original observation of the effect of interest to a critical test. Moreover, we used a sequential testing procedure (i.e., the SPRT *t* test) to increase the efficiency of data collection and applied the encoding-maintenance-retrieval MPT model to allow for an assessment of the underlying mechanisms. Finally, we decided to conduct and report our study in a registered report format to make the severity of our hypothesis tests transparent (Lakens, 2019).

Contrary to our predictions, we found that drinking alcohol immediately after learning did not significantly increase performance on a delayed subsequent memory test. In line with this finding, there was no clear evidence for a latent maintenance benefit in the alcohol condition, neither in our preregistered MPT analysis nor in an

exploratory MPT multiverse analysis. Importantly, however, these MPT analyses revealed a reliable and robust retrieval benefit in the alcohol condition. Thus, although we failed to replicate the original finding by Parker et al. (1981) using conventional measures of memory performance (i.e., cued and free recall), we did find clear-cut evidence in favor of an underlying retrieval benefit (i.e., an effect on EMR parameter *r_f*).

When evaluating our failure to successfully replicate the results by Parker et al. (1981) on the manifest memory measures of cued and free recall, several methodological differences between the original study and our replication study need to be considered. Most importantly, the results from our MPT analyses suggest that observing a strong and significant retrograde facilitation effect in behavioral measures might require a study design that maximizes the latent contributions of retrieval processes to final memory performance, while minimizing encoding and maintenance contributions. In this respect, it is important to note that Parker et al. (1981) presented their participants with pictures, while we used word pairs as learning material. Crucially, the memory-system dependent forgetting hypothesis (Hardt et al., 2013; Kuhlmann et al., 2021) suggests that hippocampus-dependent memory traces should be less susceptible to retroactive interference than extra-hippocampally represented memories. Thus, the contribution of increased retrievability as a result of reduced retroactive interference might be less pronounced when using paired associates such as word pairs as learning material. Indeed, most other studies reported in the literature used lists of single items instead of paired associates as learning material to demonstrate retrograde facilitation by alcohol (e.g., novel words, Carlyle et al., 2017). On the other hand, Parker et al. (1981) found a significant retrograde facilitation effect on recognition performance, although recognition performance should

not or only minimally rely on the retrievability of information from memory.

In contrast to our replication study and the original Parker et al. (1981) study, most studies in the literature used a retention interval between alcohol versus placebo administration and final memory test that was considerably longer than 7 h, ranging between 16 and 48 h (Bruce & Pihl, 1997; Carlyle et al., 2017; Lamberty et al., 1990; Mueller et al., 1983; Parker et al., 1980, study 2; Weafer et al., 2016a, 2016b). As a consequence, these studies could not control for memory-relevant behaviors during the retention interval and necessarily included at least one night of sleep. Against the backdrop of our replication failure, the possibility that the strength of the retrograde facilitation effect was overestimated in these studies due to an effect of acute alcohol intoxication on the sleep architecture – and, thereby, on sleep-induced memory consolidation, potentially benefitting both maintenance and retrieval processes – cannot be dismissed.

Certain limitations of our current study should also be recognized. First, in line with common conventions (cf. Cohen, 1988), we realized a (preregistered) statistical power of $1 - \beta = .80$ and a minimum effect of interest of $d = .50$ for our SPRT t tests of Hypotheses 1 and 2. Thus, compared to false-positive findings (probability $\alpha = .05$), the probability of false-negative findings ($\beta = .20$) was relatively high in our study, especially if the true d happened to be less than $.50$. However, a higher statistical power (e.g., $1 - \beta = .95$) for smaller effect sizes (e.g., $d = .20$) would have increased the expected sample size considerably. Thus, we believe to have found a reasonable compromise between sufficient statistical power on the one hand and feasibility of a very resource-intensive study procedure on the other hand.

Second, both SPRT t tests terminated rather quickly such that a decision for both Hypotheses 1 and 2 could be made after only $N = 18$ subjects had participated in the study. This outcome nicely highlights the superiority of sequential probability ratio tests in terms of efficiency (cf. Erdfelder & Schnuerch, 2021; Schnuerch & Erdfelder, 2020; Schnuerch et al., 2022). However, in our specific case, the second stopping criterion required for the MPT analyses ($n \geq 30$ in both conditions after exclusion of participants under- or overperforming in the immediate cued recall) enforced continuation of data collection until a total of $N = 93$ subjects had participated. Thus, our statistical decisions regarding Hypotheses 1 and 2 were based on a fraction of only 19% of the full sample. Although this procedure is statistically valid (i.e., it satisfies the preregistered statistical error probabilities α and β) and empirically in line with Bayesian t test outcomes, we understand that basing hypothesis tests on such a small percentage of the full sample might seem unsatisfactory. A conceivable alternative to our preregistered strategy would

have been to start the SPRT t tests only when the MPT stopping criterion had already been met. However, although avoiding the aforementioned problem, the efficiency of the data collection process would have suffered from such an approach. In our view, this shows that best practice recommendations of how to optimally combine sequential testing procedures such as the SPRT t test with additional stopping rules are needed.

Third, measured peak BACs in the alcohol condition were somewhat lower than anticipated. While the alcohol doses of 0.51 g/kg for women and 0.59 g/kg for men were expected to result in peak BACs of about 0.60‰, the observed mean peak BAC was $M = 0.43$ ‰. Two factors might be considered to explain this discrepancy. First, peak BACs were most likely reached before or after our measurements for almost all participants, such that our measures underestimate the true peak BACs. Second, alcohol administration was based on the self-reported body weights of participants. It seems reasonable to assume that participants either underestimated and/or knowingly understated their true body weight due to reasons of social desirability. Crucially, the mean peak BAC measured in our study was still higher than in the medium (0.5 ml/kg) alcohol condition of Parker et al. (1981; 0.34‰) for which these authors reported a significant retrograde facilitation effect.

Despite these limitations, our results clearly cast doubt on the reliability, strength, and generalizability of the retrograde facilitation effect in behavioral memory measures. Importantly, the effect is viewed by some authors as a central piece of evidence for the theoretical claim that memory consolidation processes are not limited to periods of sleep, but rather occur whenever retroactive interference is reduced, such as during acute alcohol intoxication (Mednick et al., 2011; Wixted, 2004, 2010). Our lack of evidence for alcohol effects on the probability of maintaining paired associates in memory suggests that such a theoretical model of episodic memory should be treated with caution. Instead, we found that postencoding alcohol administration had a beneficial effect on the probability of retrieving word pairs from memory during free recall. This result is in line with the interference hypothesis, which suggests that retrograde facilitation emerges as a direct consequence of reduced retroactive interference. As such, our finding is more in line with theoretical models that attribute episodic memory improvements (e.g., sleep-induced retrograde facilitation; Rasch & Born, 2013) to retrieval benefits resulting from increased temporal distinctiveness of memory traces (Brown et al., 2007; Ecker et al., 2015).

In sum, our mixed pattern of results best resonates with the idea that alcohol-induced retrograde facilitation really exists but is limited to a retrieval benefit caused by retroactive interference reduction. This idea nicely accounts

not only for our confirmation of Hypothesis 3 but also for our failures to confirm Hypotheses 1 and 4 (as both hypotheses refer to memory measures more sensitive to storage than to retrieval from memory). From this theoretical perspective, the only inconsistent outcome is our rejection of Hypothesis 2 concerning free recall. A possible explanation is that overall free recall performance is a less pure measure of retrieval capacity than EMR parameter r_f , which specifically captures retrieval of stored word pairs in free recall.

In terms of future research, more replication studies are needed to specify the exact conditions under which alcohol-induced retrograde facilitation can or cannot be observed. These studies should explicitly consider the role of the learning material (item vs. associative memory) and of sleep and other memory-relevant activities during the retention interval. This implies that more studies are needed that adopt the basic procedure used by Parker et al. (1981) to effectively control behaviors of participants during the retention interval. To make such resource-intensive studies feasible, they should be designed to amplify underlying retrieval differences between conditions (see above), and methodological and statistical innovations such as sequential testing procedures should be embraced. Additionally, as shown by the results of our study, the adoption of MPT modeling for hypothesis generation and testing can be expected to extend possible insights beyond the scope of more conventional analysis strategies. An a priori power analysis using multiTree (Moshagen, 2010) based on the parameter estimates from our pre-registered MPT analysis suggests a minimum total sample size of $N = 66$ across conditions (with 40 word pairs per participant) for future EMR studies that test alcohol versus placebo differences in maintenance and/or retrieval probabilities (condition difference in m or r_f under $H_1 = .10$, $\alpha = .05$, $1 - \beta = .95$).

Alcohol-induced retrograde facilitation is a fascinating psychological phenomenon with potentially far-reaching theoretical and practical implications. However, our results show that more research is clearly needed to provide a solid empirical basis for further discussions of such implications. We hope that future research will build on the methodological and statistical innovations applied in this study to arrive at a deeper understanding of the moderators and mediators of retrograde facilitation by alcohol.

Electronic Supplementary Material

The electronic supplementary material is available with the online version of the article at <https://doi.org/10.1027/1618-3169/a000569>

ESM 1. This text file provides an illustration of the EMR model and the corresponding model equations.

ESM 2. This text file details the calculation of alcohol doses as a function of gender and body weight.

ESM 3. This text file provides a complete overview of the learning material.

ESM 4. This R script contains all analysis steps to obtain the reported results (except MPT results for aggregated data).

ESM 5. This text file depicts the developments of the log-likelihood ratios of both SPRT t test.

ESM 6. This multiTree file contains all analysis steps to obtain the reported MPT results for aggregated data.

ESM 7. This text file contains the results of the MPT multiverse analysis.

References

- Babor, T. F., Higgins-Biddle, J. C., Saunders, J. B., & Monteiro, M. G. (2001). *AUDIT: The alcohol use disorders identification test: Guidelines for use in primary care* (2nd ed.). World Health Organization.
- Batchelder, W. H., & Riefer, D. M. (1999). Theoretical and empirical review of multinomial process tree modeling. *Psychonomic Bulletin & Review*, 6(1), 57–86. <https://doi.org/10.3758/BF03210812>
- Brown, G. D. A., Neath, I., & Chater, N. (2007). A temporal ratio model of memory. *Psychological Review*, 114(3), 539–576. <https://doi.org/10.1037/0033-295X.114.3.539>
- Bruce, K. R., & Pihl, R. O. (1997). Forget “drinking to forget”: Enhanced consolidation of emotionally charged memory by alcohol. *Experimental and Clinical Psychopharmacology*, 5(3), 242–250. <https://doi.org/10.1037/1064-1297.5.3.242>
- Carlyle, M., Dumay, N., Roberts, K., McAndrew, A., Stevens, T., Lawn, W., & Morgan, C. J. A. (2017). Improved memory for information learnt before alcohol use in social drinkers tested in a naturalistic setting. *Scientific Reports*, 7(1), Article 6213. <https://doi.org/10.1038/s41598-017-06305-w>
- Chan, J. K. M., Trinder, J., Andrewes, H. E., Colrain, I. M., & Nicholas, C. L. (2013). The acute effects of alcohol on sleep architecture in late adolescence. *Alcoholism: Clinical and Experimental Research*, 37(10), 1720–1728. <https://doi.org/10.1111/acer.12141>
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Routledge. <https://doi.org/10.4324/9780203771587>
- Ebrahim, I. O., Shapiro, C. M., Williams, A. J., & Fenwick, P. B. (2013). Alcohol and sleep I: Effects on normal sleep. *Alcoholism: Clinical and Experimental Research*, 37(4), 539–549. <https://doi.org/10.1111/acer.12006>
- Ecker, U. K. H., Brown, G. D. A., & Lewandowsky, S. (2015). Memory without consolidation: Temporal distinctiveness explains retroactive interference. *Cognitive Science*, 39(7), 1570–1593. <https://doi.org/10.1111/cogs.12214>
- Erdfelder, E., Auer, T.-S., Hilbig, B. E., Aßfalg, A., Moshagen, M., & Nadarevic, L. (2009). Multinomial processing tree models: A review of the literature. *Zeitschrift für Psychologie/Journal of Psychology*, 217(3), 108–124. <https://doi.org/10.1027/0044-3409.217.3.108>
- Erdfelder, E., & Schnuerch, M. (2021). On the efficiency of the independent segments procedure: A direct comparison with sequential probability ratio tests. *Psychological Methods*, 26(4), 501–506. <https://doi.org/10.1037/met0000404>

- 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>
- Fillmore, M. T. (2001). Cognitive preoccupation with alcohol and binge drinking in college students: Alcohol-induced priming of the motivation to drink. *Psychology of Addictive Behaviors*, 15(4), 325–332. <https://doi.org/10.1037/0893-164X.15.4.325>
- Gelman, A., & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, 7(4), 457–472. <https://doi.org/10.1214/ss/1177011136>
- Goodwin, D. W., Powell, B., Bremer, D., Hoine, H., & Stern, J. (1969). Alcohol and recall: State-dependent effects in man. *Science*, 163(3873), 1358–1360. <https://doi.org/10.1126/science.163.3873.1358>
- Hager, W., & Hasselhorn, M. (1994). *Handbuch deutschsprachiger Wortnormen [Handbook of German word norms]*. Hogrefe.
- Hardt, O., Nader, K., & Nadel, L. (2013). Decay happens: The role of active forgetting in memory. *Trends in Cognitive Sciences*, 17(3), 111–120. <https://doi.org/10.1016/j.tics.2013.01.001>
- Heck, D. W., Arnold, N. R., & Arnold, D. (2018). TreeBUGS: An R package for hierarchical multinomial-processing-tree modeling. *Behavior Research Methods*, 50(1), 264–284. <https://doi.org/10.3758/s13428-017-0869-7>
- Hu, X., & Batchelder, W. H. (1994). The statistical analysis of general processing tree models with the EM algorithm. *Psychometrika*, 59(1), 21–47. <https://doi.org/10.1007/BF02294263>
- Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, 90(430), 773–795. <https://doi.org/10.1080/01621459.1995.10476572>
- Keane, T. M., Lisman, S. A., & Kruzter, J. (1980). Alcoholic beverages and their placebos: An empirical evaluation of expectancies. *Addictive Behaviors*, 5(4), 313–328. [https://doi.org/10.1016/0306-4603\(80\)90005-2](https://doi.org/10.1016/0306-4603(80)90005-2)
- Klauer, K. C. (2010). Hierarchical multinomial processing tree models: A latent-trait approach. *Psychometrika*, 75(1), 70–98. <https://doi.org/10.1007/s11336-009-9141-0>
- Knowles, S. K. Z., & Duka, T. (2004). Does alcohol affect memory for emotional and non-emotional experiences in different ways?. *Behavioural Pharmacology*, 15(2), 111–121. <https://doi.org/10.1097/00008877-200403000-00003>
- Kuhlmann, B. G., Brubaker, M. S., Pfeiffer, T., & Naveh-Benjamin, M. (2021). Longer resistance of associative versus item memory to interference-based forgetting, even in older adults. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 47(3), 422–438. <https://doi.org/10.1037/xlm0000963>
- Küpper-Tetzl, C. E., & Erdfelder, E. (2012). Encoding, maintenance, and retrieval processes in the lag effect: A multinomial processing tree analysis. *Memory*, 20(1), 37–47. <https://doi.org/10.1080/09658211.2011.631550>
- Lakens, D. (2019). The value of preregistration for psychological science: A conceptual analysis. *Japanese Psychological Review*, 62(3), 221–230. https://doi.org/10.24602/sjpr.62.3_221
- Lamberty, G. J., Beckwith, B. E., Petros, T. V., & Ross, A. R. (1990). Posttrial treatment with ethanol enhances recall of prose narratives. *Physiology & Behavior*, 48(5), 653–658. [https://doi.org/10.1016/0031-9384\(90\)90206-J](https://doi.org/10.1016/0031-9384(90)90206-J)
- Lohnas, L. J., Polyn, S. M., & Kahana, M. J. (2015). Expanding the scope of memory search: Modeling intralist and interlist effects in free recall. *Psychological Review*, 122(2), 337–363. <https://doi.org/10.1037/a0039036>
- Mann, K., Batra, A., Hoch, E., & die Leitliniengruppe. (2017). S3-Leitlinie “Screening, Diagnose und Behandlung alkoholbezogener Störungen”: Kurzfassung [S3 guideline “screening, diagnosis, and treatment of alcohol-related disorders”: Summary]. *SUCHT*, 63(1), 7–24. <https://doi.org/10.1024/0939-5911/a000464>
- Mann, R. E., Cho-Young, J., & Vogel-Sprott, M. (1984). Retrograde enhancement by alcohol of delayed free recall performance. *Pharmacology Biochemistry and Behavior*, 20(4), 639–642. [https://doi.org/10.1016/0091-3057\(84\)90317-4](https://doi.org/10.1016/0091-3057(84)90317-4)
- Manthey, J., Shield, K. D., Rylett, M., Hasan, O. S. M., Probst, C., & Rehm, J. (2019). Global alcohol exposure between 1990 and 2017 and forecasts until 2030: A modelling study. *The Lancet*, 393(10190), 2493–2502. [https://doi.org/10.1016/S0140-6736\(18\)32744-2](https://doi.org/10.1016/S0140-6736(18)32744-2)
- Mednick, S. C., Cai, D. J., Shuman, T., Anagnostaras, S., & Wixted, J. T. (2011). An opportunistic theory of cellular and systems consolidation. *Trends in Neurosciences*, 34(10), 504–514. <https://doi.org/10.1016/j.tins.2011.06.003>
- Mintzer, M. Z. (2007). The acute effects of alcohol on memory: A review of laboratory studies in healthy adults. *International Journal on Disability and Human Development*, 6(4), 397–403. <https://doi.org/10.1515/IJDHD.2007.6.4.397>
- Morey, R. D., & Rouder, J. N. (2018). *BayesFactor: Computation of Bayes factors for common designs*. R package version 0.9.12-4.2. <https://CRAN.R-project.org/package=BayesFactor>
- Moshagen, M. (2010). multiTree: A computer program for the analysis of multinomial processing tree models. *Behavior Research Methods*, 42(1), 42–54. <https://doi.org/10.3758/BRM.42.1.42>
- Mueller, C. W., Lisman, S. A., & Spear, N. E. (1983). Alcohol enhancement of human memory: Tests of consolidation and interference hypotheses. *Psychopharmacology*, 80(3), 226–230. <https://doi.org/10.1007/BF00436158>
- Parker, E. S., Birnbaum, I. M., Weingartner, H., Hartley, J. T., Stillman, R. C., & Wyatt, R. J. (1980). Retrograde enhancement of human memory with alcohol. *Psychopharmacology*, 69(2), 219–222. <https://doi.org/10.1007/BF00427653>
- Parker, E. S., Morihisa, J. M., Wyatt, R. J., Schwartz, B. L., Weingartner, H., & Stillman, R. C. (1981). The alcohol facilitation effect on memory: A dose-response study. *Psychopharmacology*, 74(1), 88–92. <https://doi.org/10.1007/BF00431763>
- Quevedo Pütter, J., Erdfelder, E., & Kieslich, P. J. (2020). *Registered report protocol preregistration: Does alcohol consumption after learning really improve memory? And if so, why?* <https://osf.io/dk8fj>
- Rasch, B., & Born, J. (2013). About sleep’s role in memory. *Physiological Reviews*, 93, 681–766. <https://doi.org/10.1152/physrev.00032.2012>
- Roehrs, T., & Roth, T. (2001). Sleep, sleepiness, and alcohol use. *Alcohol Research & Health*, 25(2), 101–109. <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6707127/>
- Rouder, J. N., & Batchelder, W. H. (1998). Multinomial models for measuring storage and retrieval processes in paired associate learning. In C. E. Dowling, F. S. Roberts, & P. Theuns (Eds.), *Recent progress in mathematical psychology: Psychophysics, knowledge, representation, cognition, and measurement* (pp. 195–225). Lawrence Erlbaum Associates Publishers.
- Saunders, J. B., Aasland, O. G., Babor, T. F., de la Fuente, J. R., & Grant, M. (1993). Development of the Alcohol Use Disorders Identification Test (AUDIT): WHO collaborative project on early detection of persons with harmful alcohol consumption-II. *Addiction*, 88(6), 791–804. <https://doi.org/10.1111/j.1360-0443.1993.tb02093.x>
- Schmidt, O., Erdfelder, E., & Heck, D. W. (2023). How to develop, test, and extend multinomial processing tree models: A tutorial. *Psychological Methods*. Advance online publication. <https://doi.org/10.1037/met0000561>
- Schnuerch, M., & Erdfelder, E. (2020). Controlling decision errors with minimal costs: The sequential probability ratio t test. *Psychological Methods*, 25(2), 206–226. <https://doi.org/10.1037/met0000234>
- Schnuerch, M., Heck, D. W., & Erdfelder, E. (2022). Waldian t tests: Sequential Bayesian t tests with controlled error probabilities. *Psychological Methods*. Advance online publication. <https://doi.org/10.1037/met0000492>

- Steinhilber, M., Schnuerch, M., & Schubert, A.-L. (2021). *sprtt: Sequential Probability Ratio Tests: Using t-Statistic*. R package version 0.1.0. <https://CRAN.R-project.org/package=sprtt>
- Thierauf, A., Kempf, J., Eschbach, J., Auwärter, V., Weinmann, W., & Gnann, H. (2013). A case of a distinct difference between the measured blood ethanol concentration and the concentration estimated by Widmark's equation. *Medicine, Science and the Law*, 53(2), 96–99. <https://doi.org/10.1258/msl.2012.012038>
- Tucker, M. A., Hirota, Y., Wamsley, E. J., Lau, H., Chaklader, A., & Fishbein, W. (2006). A daytime nap containing solely non-REM sleep enhances declarative but not procedural memory. *Neurobiology of Learning and Memory*, 86(2), 241–247. <https://doi.org/10.1016/j.nlm.2006.03.005>
- Tulving, E., & Psotka, J. (1971). Retroactive inhibition in free recall: Inaccessibility of information available in the memory store. *Journal of Experimental Psychology*, 87(1), 1–8. <https://doi.org/10.1037/h0030185>
- Tyson, P. D., & Schirmuly, M. (1994). Memory enhancement after drinking ethanol: Consolidation, interference, or response bias? *Physiology & Behavior*, 56(5), 933–937. [https://doi.org/10.1016/0031-9384\(94\)90326-3](https://doi.org/10.1016/0031-9384(94)90326-3)
- Weafer, J., Gallo, D. A., & De Wit, H. (2016a). Acute effects of alcohol on encoding and consolidation of memory for emotional stimuli. *Journal of Studies on Alcohol and Drugs*, 77(1), 86–94. <https://doi.org/10.15288/jsad.2016.77.86>
- Weafer, J., Gallo, D. A., & De Wit, H. (2016b). Effect of alcohol on encoding and consolidation of memory for alcohol-related images. *Alcoholism: Clinical and Experimental Research*, 40(7), 1540–1547. <https://doi.org/10.1111/acer.13103>
- Wechsler, H., Dowdall, G. W., Davenport, A., & Rimm, E. B. (1995). A gender-specific measure of binge drinking among college students. *American Journal of Public Health*, 85(7), 982–985. <https://doi.org/10.2105/AJPH.85.7.982>
- White, N. M. (1996). Addictive drugs as reinforcers: Multiple partial actions on memory systems. *Addiction*, 91(7), 921–950. <https://doi.org/10.1046/j.1360-0443.1996.9179212.x>
- Wixted, J. T. (2004). The psychology and neuroscience of forgetting. *Annual Review of Psychology*, 55, 235–269. <https://doi.org/10.1146/annurev.psych.55.090902.141555>
- Wixted, J. T. (2010). The role of retroactive interference and consolidation in everyday forgetting. In S. Della Sala (Ed.), *Forgetting* (pp. 285–312). Psychology Press.

History

Received December 31, 2019
 Revision received September 29, 2022
 Accepted December 12, 2022
 Published online February 21, 2023

Acknowledgments

We would like to thank Pascal J. Kieslich for his help with preparing the study, and Lena Bizer, Annalena Loose, Linus Quevedo Pütter, Pia Quevedo Pütter, Christina Sarafoglou, and Julian Ziegler for their help with data collection.

Publication Ethics

The research protocol was approved by the ethics committee of the University of Mannheim.



Open Data

The Stage-1 manuscript of this registered report (<https://doi.org/10.17605/OSF.IO/DK8FJ>; Quevedo Pütter et al., 2020), the materials (<https://doi.org/10.17605/OSF.IO/8E9PW>), and the data (<https://doi.org/10.17605/OSF.IO/2K4J7>) are publicly available on the Open Science Framework (OSF). Additional information needed to reproduce all reported results is provided in the electronic supplemental material (ESM).

Funding

This research was funded by the Deutsche Forschungsgemeinschaft (DFG), Grant 2277, Research Training Group “Statistical Modeling in Psychology” (SMiP). Open access publication enabled by University of Mannheim.


ORCID


J. Quevedo Pütter
 <https://orcid.org/0000-0002-7340-9937>
 E. Erdfelder
 <https://orcid.org/0000-0003-1032-3981>


Edgar Erdfelder

Cognition and Individual Differences Lab
 University of Mannheim
 68131 Mannheim
 Germany
erdfelder@uni-mannheim.de

Author Note

Julian Quevedo Pütter  <https://orcid.org/0000-0002-7340-9937>

Selina Dahler  <https://orcid.org/0009-0004-0245-8506>

Edgar Erdfelder  <https://orcid.org/0000-0003-1032-3981>

This manuscript was written in R Markdown using the *papaja* package (Aust & Barth, 2023). A fully reproducible manuscript version (including all data analysis steps), the preregistrations of all three experiments, the corresponding data sets, and the experimental software are available at osf.io/fsb76/?view_only=8921383d9c054233a13b0642c5aa92c9.

This research was funded by the Deutsche Forschungsgemeinschaft (DFG), grant 2277, Research Training Group “Statistical Modeling in Psychology” (SMiP). Portions of this research were presented at the 2022 IOPS-SMiP Summer Conference in Leuven, Belgium, and the 2023 Conference of Experimental Psychologists (TeaP) in Trier, Germany. The initial version of this manuscript was included in the cumulative dissertation by Julian Quevedo Pütter. The authors have no conflicts of interest to declare.

Author contributions: Julian Quevedo Pütter: Conceptualization (lead), Data curation, Formal analysis, Investigation, Methodology (lead), Project administration (lead), Software, Validation, Visualization, Writing—original draft, Writing—review and editing; Selina Dahler: Conceptualization (supporting), Investigation, Project administration (supporting), Validation, Writing—review and editing; Edgar Erdfelder: Conceptualization (supporting), Funding acquisition, Methodology (supporting), Resources, Supervision, Writing—review and editing.

28 We would like to thank Ullrich Ecker for providing the raw data of Ecker, Brown, et
29 al. (2015) that was used for the reanalysis reported in the Introduction section and as part
30 of the analysis of Experiment 1.

31 Correspondence concerning this article should be addressed to Julian Quevedo Pütter,
32 Department of Psychology, University of Mannheim, B6 30-32, 68159 Mannheim, Germany.
33 E-mail: julian@quevedo.de

34

Abstract

35 Retroactive interference in episodic memory is assumed to follow a temporal gradient.
36 Specifically, delaying an interpolated learning phase has been shown to benefit subsequent
37 memory performance with respect to an original learning phase. Much of the previous
38 research on the temporal gradient of retroactive interference (TGRI) has been affected by
39 methodological weaknesses which may explain its rather low replication rate (Wixted, 2004).
40 Moreover, the TGRI has been interpreted as key evidence in favor of opportunistic
41 consolidation theory (Mednick et al., 2011). This interpretation is challenged by an
42 alternative theoretical account, namely, temporal distinctiveness theory (Brown et al., 2007).
43 In the current research, we aimed at (a) establishing the replicability of the TGRI, and (b)
44 deciding between both theoretical accounts by means of a storage-retrieval multinomial
45 processing tree (MPT) model. In three preregistered experiments (total $N = 397$),
46 participants were asked to learn and retrieve word lists across multiple trials. Crucially, an
47 interpolated item list was presented either rather early or rather late during the 5-min
48 retention interval. We found a robust TGRI in recall but not in recognition across
49 experiments. Our MPT results consistently show that the TGRI is purely retrieval-driven,
50 without any storage contribution. This is in line with temporal distinctiveness theory but in
51 conflict with the opportunistic consolidation account. Surprisingly, the similarity of original
52 and interpolated learning materials had no effect on any memory measure. Thus, theories of
53 retroactive interference might dispense with a consolidation mechanism and should
54 reconsider the importance of similarity.

55 *Public Significance Statement:* This research shows that increasing the time interval
56 between two unrelated learning units benefits subsequent memory for the material from the
57 first learning unit. A longer time interval between units does not affect the storage of the
58 material in memory but rather benefits the ease of retrieving it from memory. This benefit is
59 independent of the similarity of the materials presented during the two learning units.

60 *Keywords:* episodic memory, retroactive interference, consolidation, temporal
61 distinctiveness, multinomial processing tree modeling

62 Word count: 15082

63 **Opportunistic Consolidation or Temporal Distinctiveness? Retrieval, Not**
64 **Storage, Drives the Temporal Gradient of Retroactive Interference in Episodic**
65 **Memory**

66 Retroactive interference is considered to be a major source of everyday forgetting in
67 human episodic memory: Whenever encoding of a new piece of information is followed by
68 encoding of some unrelated information, later recall of the original information can be
69 expected to be impaired. Indeed, human beings engage in intentional or incidental learning
70 activities almost continuously; they “never stop making memories” (Wixted, 2010, p. 290).
71 Thus, virtually all newly encoded memories will have to endure some more or less severe
72 influence of retroactive interference.

73 Consequently, extensive research efforts have been devoted towards investigating
74 memory benefits resulting from at least temporarily minimizing retroactive interference.
75 Whereas the facilitating effect of sleep after learning is a well-established phenomenon (for a
76 review and meta-analysis, see Berres & Erdfelder, 2021), the exact conditions under which
77 retroactive interference minimization results in a memory benefit during wakefulness remain
78 controversial. Many studies have found that a short period of post-encoding waking rest can
79 benefit subsequent memory performance compared to a cognitively demanding distractor
80 task even if original and interpolated materials are dissimilar (see Wamsley, 2019). Indeed,
81 waking rest has been shown to facilitate not only declarative (e.g., Martini et al., 2020) and
82 procedural memory (e.g., Humiston & Wamsley, 2018), but even insight into complex
83 problem solving (Craig et al., 2018). Whereas these findings ascribe waking rest a role in
84 post-encoding processing similar to sleep, other studies have failed to replicate such effects
85 (see Martini & Sachse, 2020). For example, Varma et al. (2017) compared the effect of
86 waking rest against different variants of a cognitively demanding n-back task on recall and
87 recognition across six experiments (total $N = 176$). In none of their experiments were they
88 able to detect a difference between conditions. Thus, it remains to be determined under

89 exactly which conditions minimizing retroactive interference during wakefulness results in
90 improved memory.

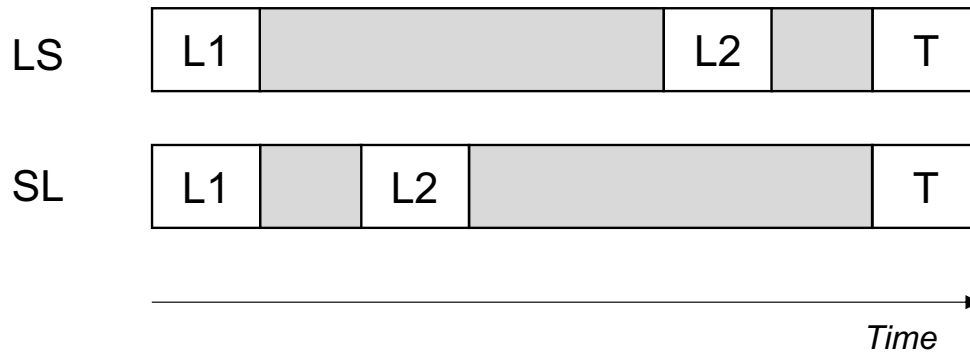
91 Instead of attempting to *minimize* retroactive interference across the entire retention
92 interval, an alternative approach in the field has been to merely *delay* the onset of
93 interpolated learning or a distractor task. It might be argued that such an approach
94 increases external validity by more explicitly acknowledging the ubiquity of encoding
95 demands in everyday life. Beginning with the pioneering work of Müller and Pilzecker
96 (1900), researchers have employed a study design that involves learning of some original
97 material (L1), followed by some interpolated learning (L2), and a final memory test for the
98 original material (T). Crucially, the temporal position of the L2 learning phase within the
99 L1-T retention interval is manipulated: In “Long-Short” (LS) conditions, the L1-L2 interval
100 is relatively long, whereas the L2-T interval is relatively short. In contrast, in “Short-Long”
101 (SL) conditions, the L1-L2 interval is relatively short, whereas the L2-T interval is relatively
102 long. This study design is illustrated in Figure 1.

103 Past research using such a study design or extensions thereof suggests that retroactive
104 interference might follow a temporal gradient, that is, final L1 memory performance benefits
105 from delaying L2 learning. In other words, memory performances can generally be expected
106 to be better in LS compared to SL conditions. Crucially, empirical demonstrations of the
107 temporal gradient of retroactive interference (TGRI) have been similarly ambiguous as for
108 the waking rest effect (for a review and critical discussion, see Wixted, 2004). Thus, the
109 robustness of the effect remains to be determined. One of our main goals in the current
110 research was to establish the replicability of the TGRI within a paradigm proposed by Ecker,
111 Brown, et al. (2015) across different experimental settings (online, laboratory), and memory
112 tests (recall, recognition).

113 Notwithstanding its ambiguous empirical foundation, along with the waking rest
114 effect, the TGRI is seen by many researchers as a central piece of evidence in favor of an

Figure 1

Study Design Required to Demonstrate the Temporal Gradient of Retroactive Interference



Note. LS = "Long-Short", SL = "Short-Long" condition. L1 = first learning phase, L2 = second learning phase, T = memory test for L1 items. In the LS condition, the L1-L2 interval is relatively long, whereas the L2-T interval is relatively short; in the SL condition, the L1-L2 interval is relatively short, whereas the L2-T interval is relatively long. Illustration adapted from Ecker, Brown, et al. (2015).

115 opportunistic theory of episodic memory consolidation (see Dewar et al., 2007; Mednick et
116 al., 2011; Wixted, 2004). According to this theory, the encoding of new information and the
117 consolidation of previously encoded information are mutually exclusive. More specifically,
118 hippocampal resources are assumed to be limited and to be used for either encoding or
119 consolidation. Thus, whenever encoding demands are temporarily reduced, hippocampal
120 resources are opportunistically reallocated to critical consolidation processes. When
121 comparing LS and SL conditions, it is assumed that L1 memories benefit from delaying L2
122 learning because there is more time for L1 consolidation before L2 learning interrupts the
123 consolidation process. Since successful consolidation is thought to determine the long-term
124 fate of memories, it has been claimed that this daytime consolidation mechanism is a major
125 contributor to everyday memory (Wixted, 2004, 2010).

126 Such an optimistic interpretation is challenged not only by the rather low reliability
127 of the TGRI, but also by the existence of at least one alternative theoretical account:
128 Temporal distinctiveness theory (Brown et al., 2007; Ecker, Brown, et al., 2015) provides a
129 parsimonious explanation of the TGRI that does not rely on a consolidation mechanism.
130 Instead, the basic assumption underlying distinctiveness theory is that retrieval of an item
131 from memory is determined by that item’s discriminability from other memory
132 representations within a latent psychological space. This space may be defined by various
133 relevant dimensions: According to the *temporal* distinctiveness theory, one such dimension is
134 a temporal one, whereby the discriminability – and, as a consequence, retrievability – of a
135 memory is thought to be determined by its temporal isolation from other memories encoded
136 before or afterwards. Thus, a memory representation that is more isolated on this temporal
137 dimension should be more easily retrievable than a memory representation that lies in close
138 temporal proximity to other memories. It follows that delaying L2 learning will increase the
139 temporal distinctiveness of L1 memories, leading to overall better final L1 memory
140 performance. Importantly, extending such a purely *temporal* distinctiveness theory to a
141 *generalized* distinctiveness theory suggests that any number of L1 and L2 characteristics –

142 such as, for example, their semantic overlap – may also contribute to the overall
143 discriminability and retrievability of L1 items.

144 Unfortunately, at first glance, the explanations of opportunistic consolidation and
145 temporal distinctiveness theory with respect to the TGRI seem inextricable since both
146 predict that delaying L2 learning will facilitate final L1 memory performance. For example,
147 Mercer (2015) demonstrated the TGRI in a convincing manner but had to conclude that his
148 experiment “cannot disentangle consolidation and distinctiveness-based accounts.” Future
149 studies would thus need “an ingenious design to fully extricate the predictions of these two
150 accounts” (p. 134). Rather than relying on an “ingenious” study design, we believe one
151 solution to finally disentangle both proposed mechanisms lies in using an appropriate
152 cognitive measurement model that disentangles the relevant underlying processes. Our
153 second main goal in the current research was to precisely differentiate between opportunistic
154 consolidation and temporal distinctiveness contributions to the TGRI. To achieve both goals,
155 we adapted an established storage-retrieval multinomial processing tree (MPT) model to the
156 TGRI paradigm.

157 The issues of replicability and differentiation of explanatory accounts are highly
158 critical to the investigation of the TGRI in episodic memory. In the following, we outline in
159 more detail how we aimed to address both issues simultaneously in our current research.

160 **Resolving Replicability Issues**

161 Müller and Pilzecker (1900) were the first to predict and empirically determine the
162 existence of the TGRI in episodic memory. In their experiment, a single participant was
163 repeatedly asked to study lists of syllable pairs. In the LS condition, the L1-L2 interval was
164 6 minutes long, whereas in the SL condition, it only lasted about 17 seconds (i.e., the time
165 necessary to switch the apparatus from L1 to L2 presentation). The L1-T retention interval
166 had a total duration of about 90 minutes in all trials. The authors observed that L1 cued

167 recall rates were higher and response times shorter in the LS than in the SL condition.

168 However, subsequent research on the TGRI throughout the 20th century has led to a
169 rather mixed body of evidence. Whereas some studies found an effect (e.g., Landauer, 1974;
170 Skaggs, 1925), others failed to replicate it (e.g., McGeoch, 1933; Robinson, 1920; Wickelgren,
171 1974). Based on these conflicting results, Wickelgren (1977) concluded that the hypothesis of
172 a TGRI in episodic memory must be rejected.

173 In contrast, in a more nuanced review and reanalysis of the available data, Wixted
174 (2004) convincingly argued for a reinterpretation of the existing evidence. According to his
175 reasoning, past replication failures could mainly be attributed to two methodological issues.
176 First, in line with an opportunistic consolidation explanation of the effect, he suggested that
177 for the TGRI to emerge, cognitive demands between learning and test phases (i.e., during
178 the L1-L2 and the L2-T intervals) must be minimal. Otherwise, consolidation processes
179 might be inhibited throughout the entire retention interval. Thus, instead of letting
180 participants read newspaper articles during L1-L2 and L2-T intervals (see Robinson, 1920)
181 or even applying a continuous recognition paradigm associated with high encoding demands
182 during L1-L2 and L2-T intervals (see Wickelgren, 1974), experimenters should minimize
183 cognitive demands, as in case of waking rest. Notably, however, Landauer (1974) found a
184 TGRI by inducing reduced instead of minimized encoding demands during the L1-L2 and
185 L2-T intervals. More specifically, in all of his experiments, L1-L2 and L2-T intervals involved
186 learning of relatively less attention-demanding materials. His results imply that a reduction
187 of cognitive demands during L1-L2 and L2-T intervals relative to L2 learning might already
188 suffice to allow the TGRI to emerge.

189 Second, Wixted (2004) emphasized the importance of a delay between interpolated
190 L2 learning and the final L1 test. If the L2-T interval becomes too short, retrieval of L1
191 items will temporarily be inhibited. Crucially, L1 retrieval inhibition counteracts the TGRI,
192 since memory performance will be more negatively biased in the LS than in the SL condition.

193 Indeed, retrieval inhibition might even reverse the result pattern: In research reported by
194 McGeoch (1933), L2 learning occurred either immediately after L1 learning (SL) or
195 immediately before the final L1 memory test (LS). As a result, memory performance was
196 better in the SL than in the LS condition. Thus, researchers should not undercut some
197 minimum L2-T interval to exclude the possibility of retrieval inhibition.

198 More recent investigations of the TGRI can be evaluated with respect to these
199 methodological recommendations and their observed outcomes. For example, Ecker, Brown,
200 et al. (2015) conducted two experiments (total $N = 48$) that involved word lists as L1 and
201 L2 material. During the L1-L2 and L2-T phases, participants worked on a highly controlled
202 yet simple tone-detection (Experiment 1) or number reading task (Experiment 2). The study
203 design involved not only an LS and an SL condition, but additionally an LL and an SS
204 condition. Depending on the condition, the L2-T interval had a duration of either 60 seconds
205 (LS and SS) or 240 seconds (SL and LL). Thus, there was a clear difference in encoding
206 demands between L2 learning and the distractor tasks, and retrieval inhibition was
207 prevented. Interestingly, despite both experiments differing only in the respective distractor
208 task, a reanalysis of the two critical conditions LS and SL reveals a pronounced TGRI in
209 Experiment 1, $M_{LS} = 0.48$ ($SD = 0.12$), $M_{SL} = 0.38$ ($SD = 0.14$), $t(22) = 3.43$, $p = .001$,
210 Hedges' $\hat{g} = 0.70$, but not in Experiment 2, $M_{LS} = 0.35$ ($SD = 0.17$), $M_{SL} = 0.33$ ($SD =$
211 0.12), $t(24) = 0.97$, $p = .172$, Hedges' $\hat{g} = 0.16$. Since Ecker, Brown, et al. (2015) adhered to
212 both methodological recommendations by Wixted (2004), the problem might rather be a lack
213 of statistical power of our reanalysis due to its focus on only two out of four conditions.

214 Given that descriptive tendencies in line with the TGRI emerged in both experiments,
215 we chose to adapt the paradigm used by Ecker, Brown, et al. (2015) in our current research
216 using sample sizes that guarantee sufficiently high power for detecting potential LS versus SL
217 differences.

218 Disentangling Mechanisms Proposed to Underlie the TGRI

219 So far, attempts to differentiate between opportunistic consolidation and temporal
220 distinctiveness theory on a behavioral level have been scarce. Nevertheless, the TGRI has
221 often been cited as evidence in favor of opportunistic consolidation theory (e.g., Dewar et al.,
222 2007; Wixted, 2004).

223 Some indirect evidence for and against both accounts can be derived from previous
224 research on the TGRI. For example, Mercer (2015) used Icelandic-English word pairs as L1
225 material and manipulated L2 timing such that the interpolated L2 learning phase occurred
226 immediately following the L1 learning phase (SL) or after an 8-min delay of waking rest (LS).
227 Additionally, he manipulated the similarity of L1 and L2 materials, such that L2 learning
228 involved either Norwegian-English word pairs (high L1-L2 similarity) or face pairs (low
229 L1-L2 similarity). A significant TGRI for cued recall performances emerged regardless of
230 L1-L2 similarity. From an opportunistic consolidation perspective, the absence of an
231 interaction effect of L2 timing and L1-L2 similarity comes as no surprise, since any
232 cognitively demanding learning phase, regardless of its similarity to the original learning
233 phase, should inhibit consolidation. In contrast, from the perspective of a generalized
234 distinctiveness theory, the temporal dimension is just one of many latent dimensions that
235 together make up the memory space (see Ecker, Brown, et al., 2015). Thus, low L1-L2
236 similarity might have been expected to reduce the overall strength of any retroactive
237 interference effect, possibly resulting in a TGRI only for high, but not for low L1-L2
238 similarity. Therefore, the absence of any L1-L2 similarity effect in the experiment by Mercer
239 (2015) is more easily explained by opportunistic consolidation theory.

240 Additional research by Ecker, Tay, et al. (2015) supports the opposite conclusion.
241 Whereas both opportunistic consolidation and temporal distinctiveness theory predict a
242 positive effect of post-encoding retroactive interference minimization, a positive pre-encoding
243 minimization effect is only implied by temporal distinctiveness theory. In two experiments,

244 Ecker, Tay, et al. (2015) had their participants learn three word lists per trial, with final free
245 recall tests always targeting the second list of the respective trial. In a 2 x 2 design, the
246 durations of both pre- and post-encoding intervals were manipulated to be either “long” (120
247 seconds in Experiment 1, 60 seconds in Experiment 2) or “short” (15 seconds in both
248 experiments). They found that a longer pre-encoding interval and, to a lesser extent, a
249 longer post-encoding interval both facilitated final recall. This pattern of results is easily
250 explained by temporal distinctiveness theory but cannot be reconciled with opportunistic
251 consolidation theory without speculative post-hoc assumptions (see Ecker, Tay, et al., 2015).

252 Some of the most direct evidence against an opportunistic consolidation explanation
253 of the TGRI comes from research reported by Ecker, Brown, et al. (2015, see above). In a
254 model-based approach, the authors applied a specific implementation of the temporal
255 distinctiveness theory to the serial position curve data from both their experiments, the
256 SIMPLE (Scale-Invariant Memory, Perception, and Learning) model (Brown et al., 2007).
257 Their idea was to compare model fits between many different model versions, with some of
258 them implementing a consolidation mechanism. Overall, model fit indices clearly favored
259 consolidation-free model versions for both experiments, that is, the increase in model
260 complexity from adding additional consolidation parameters was too large to justify
261 relatively small improvements in model fit.

262 Critically, the underspecification of the consolidation process in the literature forced
263 Ecker, Brown, et al. (2015) to consider a large number of conceivable model
264 implementations. Indeed, the exact rate, functional form, and time-scale of episodic memory
265 consolidation remain to be determined (Ecker & Lewandowsky, 2012). Therefore, the
266 authors explored different assumptions about the exact starting point, end point, and shape
267 of the consolidation mechanism in their model implementations. This lack of more precise
268 predictions from opportunistic consolidation theory not only complicated the elaborate
269 computational modeling approach by Ecker, Brown, et al. (2015), but might also be the

270 reason why more comprehensive evidence that would allow for a definitive decision between
271 opportunistic consolidation and temporal distinctiveness is still pending.

272 Intriguingly, both theoretical accounts make identical predictions only on a surface
273 level, that is, they both predict a longer L1-L2 interval to facilitate subsequent L1 memory
274 performance. However, conventional memory measures such as recognition and recall
275 accuracy are driven by underlying storage and retrieval contributions. More specifically,
276 whenever a correct response is observed on some memory test that cannot be attributed to
277 lucky guessing, the respective memory representation must necessarily exist (i.e., it was
278 successfully *stored* in memory) and also be *retrievable* at the precise moment of the
279 recognition or recall prompt. In case of an incorrect response, however, it remains unclear
280 whether the respective memory representation does not exist (i.e., *storage* was unsuccessful)
281 or whether merely *retrieval* of an existing memory representation failed.

282 Temporal distinctiveness theory explicitly predicts that any forgetting is purely
283 retrieval-based, that is, manipulating the temporal isolation of memory representations will
284 influence their retrievability whereas storage should remain unaffected (Brown et al., 2007;
285 Ecker, Brown, et al., 2015). Fortunately, despite its overall underspecification, equally
286 precise predictions can be derived from opportunistic consolidation theory. Typically, two
287 phases of consolidation processes are distinguished, namely, synaptic (or cellular)
288 consolidation and systems consolidation (Dudai, 2004; Mednick et al., 2011). Whereas
289 systems consolidation describes the gradual integration of new information with pre-existing
290 memories and is thought to occur over a rather long time-course of several months or even
291 years (McClelland et al., 1995; Mednick et al., 2011), synaptic consolidation should be more
292 relevant for retention intervals of a few minutes. Here, immediately following initial encoding,
293 new memory traces are stabilized by strengthening synaptic connections through long-term
294 potentiation (Dudai, 2004; McGaugh, 2000). Thus, rather than affecting retrieval, synaptic
295 consolidation that occurs within the first minutes after encoding should primarily benefit

296 successful storage of new information. Indeed, a negligible role of retrieval processes has been
297 explicitly proposed in theoretical models of opportunistic consolidation before (see Dewar et
298 al., 2007).

299 Disentangling storage and retrieval contributions to memory performance thus offers
300 an exciting new possibility to finally subject both the opportunistic consolidation and the
301 temporal distinctiveness theory to a rigorous test. Whereas the opportunistic consolidation
302 theory predicts the TGRI to be driven by storage, the temporal distinctiveness theory views
303 the TGRI as a retrieval phenomenon. Importantly, both storage and retrieval might be
304 involved in producing the TGRI, that is, opportunistic consolidation and temporal
305 distinctiveness are in fact not mutually exclusive but could both contribute to the TGRI.

306 *A Storage-Retrieval MPT Approach*

307 Multinomial processing tree (MPT) models allow researchers to precisely disentangle
308 latent cognitive processes underlying observable response data (for a review of MPT model
309 applications, see Erdfelder et al., 2009; for a tutorial on MPT modeling, see Schmidt et al.,
310 2023). So-called storage-retrieval MPT models (Nadarevic, 2017; Riefer & Batchelder, 1995)
311 are tailored to experimental paradigms that involve the learning and subsequent recognition
312 or recall of previously learned items. To illustrate, a failure to recall an item in a free recall
313 test might be the result of either failed item storage or failed retrieval, and a correct
314 response in a recognition test might be the result of successful storage and retrieval or lucky
315 guessing. Thus, responses in memory tests reflect the interplay of different cognitive
316 processes that remain unobservable if not subjected to an appropriate MPT analysis.

317 For our purposes, we adapted an established storage-retrieval MPT model for the
318 recognition-then-cued-recall paradigm (Riefer & Batchelder, 1995) to a
319 free-recall-then-recognition paradigm. While maintaining the basic logic of the underlying
320 processing steps, this adapted model allows for a close replication of the paradigm proposed

321 by Ecker, Brown, et al. (2015) since it only requires an additional old-new recognition test at
322 the end of each trial. By switching the test order, we intended to avoid a situation where
323 participants have a chance to restudy target items during the recognition test for subsequent
324 free recall. We also incorporated important model adaptations proposed by Nadarevic (2017).

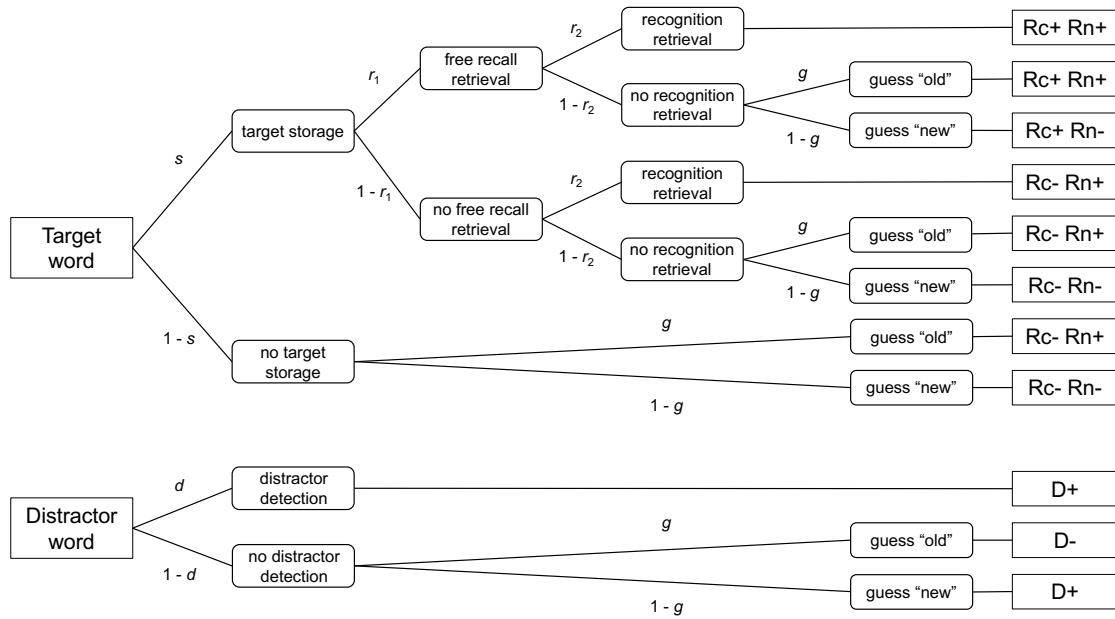
325 Such a procedure yields a 2 (recall: success vs. failure) x 2 (recognition: success
326 vs. failure) matrix of observable responses for items presented during the original L1 learning
327 phase. Our storage-retrieval MPT analyses thus aimed at explaining the probabilities of the
328 resulting four response categories: successful recall and successful recognition (“Rc+ Rn+”),
329 successful recall but unsuccessful recognition (“Rc+ Rn-”), unsuccessful recall but successful
330 recognition (“Rc- Rn+”), and unsuccessful recall and unsuccessful recognition (“Rc- Rn-”).

331 To explain the probabilities of these response categories, our storage-retrieval MPT
332 model employs four model parameters s , r_1 , r_2 , and g , reflecting successful storage,
333 successful retrieval in the free recall and recognition tests, and guessing “old” in the
334 recognition test, respectively. Specifically, an L1 item is successfully stored with probability
335 s , whereas storage is unsuccessful with probability $1 - s$. In the free recall test, a stored item
336 is successfully retrieved with probability r_1 , whereas free recall retrieval of a stored item fails
337 with probability $1 - r_1$. In the recognition test, a stored item is successfully retrieved with
338 probability r_2 , whereas recognition retrieval of a stored item fails with probability $1 - r_2$. In
339 case of recognition retrieval failure, a participant might still provide a correct answer
340 through guessing. This is implemented in the model by assuming that a participant will
341 correctly guess “old” with probability g , whereas they will incorrectly guess “new” with
342 probability $1 - g$.

343 Following Nadarevic (2017), we introduced a second tree into our model that
344 represents the processing of distractor items in the recognition test. Only two response
345 categories must be considered, namely correct (“D+”) versus incorrect classification of a
346 distractor item (“D-”). A given distractor item might be successfully detected with

Figure 2

Illustration of the Adapted Riefer and Batchelder (1995) Storage-Retrieval MPT Model



Note. MPT = multinomial processing tree, s = probability of successfully storing a target word in memory, r_1 = probability of successfully retrieving a stored target word in free recall, r_2 = probability of successfully retrieving a stored target word in recognition, g = probability of guessing "old" during recognition given no recognition retrieval or distractor detection, d = probability of detecting a distractor during recognition. Rc+ = successful target recall, Rc- = unsuccessful target recall, Rn+ = successful target recognition, Rn- = unsuccessful target recognition, D+ = correct distractor classification, D- = incorrect distractor classification.

347 probability d , whereas distractor detection might fail with probability $1 - d$. Given failed
348 distractor detection, a participant might incorrectly guess “old” with probability g , whereas
349 they will correctly guess “new” with probability $1 - g$. Following again Nadarevic (2017) by
350 linking d to other model parameters, one degree of freedom is gained from including a
351 distractor tree without introducing additional parameters. Figure 2 illustrates the complete
352 storage-retrieval MPT model involving both the target and the distractor tree.

353 By extending the original Ecker, Brown, et al. (2015) procedure to include a
354 recognition test at the end of each trial, we were able to apply the adapted Riefer and
355 Batchelder (1995) storage-retrieval MPT model to these data and disentangle storage and
356 retrieval contributions to observed memory performances.

357 **Overview of the Current Experiments**

358 In the current research, we conducted three experiments to (a) establish the
359 replicability of the TGRI within the paradigm proposed by Ecker, Brown, et al. (2015,
360 Experiment 1), and (b) precisely differentiate between opportunistic consolidation and
361 temporal distinctiveness contributions to the TGRI by means of the storage-retrieval MPT
362 model. Whereas Experiments 1 and 2 were conducted in an online setting, Experiment 3
363 took place in a more controlled lab environment.

364 ***Transparency and Openness***

365 For each experiment, we report in detail our sample size rationale, the evaluation and
366 application of preregistered exclusion criteria, our experimental manipulations, our measures,
367 and our model specifications. A reproducible version of this manuscript including the
368 analysis code, as well as the preregistrations, data, and materials from all three experiments
369 have been made publicly available at the Open Science Framework (OSF) and can be
370 accessed at osf.io/fsb76/?view_only=8921383d9c054233a13b0642c5aa92c9.

Experiment 1

371

372 Experiment 1 tested the replicability of the TGRI as found by Ecker, Brown, et al.
373 (2015, Experiment 1). Switching to an online setting required some minor changes to the
374 original material and procedure. We only included the critical LS and SL conditions from
375 the original study design.

376 To evaluate replication success, we tested the null hypothesis that the TGRI is
377 equally large in the original study by Ecker, Brown, et al. (2015) and in our replication
378 study. Hence, instead of analyzing our replication data in isolation, we obtained the original
379 data from Ecker, Brown, et al. (2015) and included them in a 2 x 2 mixed ANOVA model
380 with factors “study” (original vs. replication, between-participants) and “L2 timing” (LS
381 vs. SL, within-participants).¹ A successful replication would be indicated by a non-significant
382 between-within interaction effect of study and L2 timing on free recall accuracy, provided
383 that this ANOVA interaction test is sufficiently powered to detect replication failures with
384 high probability (see Participants section). In other words, our replication goal was not
385 merely to test for a TGRI effect of any size, but to more rigorously assess whether our
386 replication results are consistent with the original results (see Anderson & Maxwell, 2016).

387 Prior to data collection, a detailed study protocol including our hypotheses and the
388 analysis plan was uploaded to the OSF (osf.io/afxpbe).

389 Method

390 *Participants*

391 We used the software program G*Power (Faul et al., 2009) to conduct an a priori
392 power analysis for our critical ANOVA between-within interaction F test given $\alpha = .05$.
393 More precisely, we aimed at a high power of $1 - \beta = .95$ to detect a difference in effect sizes

¹ We are very grateful to Ullrich Ecker for providing the well-documented raw data of Ecker, Brown, et al. (2015).

394 of the original and the replication study that corresponds to a complete replication failure,
395 that is, a true TGRI population effect size of Cohen's $d_r = 0$ in our replication. If d_o denotes
396 the true population effect size in the original study and $d_r = 0$ the effect size representing
397 replication failure, then this implies a critical interaction effect of size $f = (d_o - d_r)/4 = d_o/4$
398 (Cohen, 1988). We made use of Hedges' bias-corrected effect size measure \hat{g}_o to estimate d_o
399 (see, e.g., Lakens, 2013). Based on the LS versus SL condition data observed in the original
400 study, we obtained Hedges' $\hat{g}_o = 0.698$, resulting in a to-be-detected interaction effect size of
401 Cohen's $f = 0.698/4 = 0.174$. Since we were only interested in detecting a replication effect
402 size that is significantly *smaller* (not larger) than the original effect size (i.e., $H_1 : d_r < d_o$),
403 we preregistered a one-tailed interaction test with $\alpha = .05$ that corresponds to a standard
404 two-tailed F test with $\alpha = .10$, provided that the observed TGRI effect size is in the
405 predicted direction (i.e., $\hat{g}_r < 0.698$). Moreover, since the observed correlation between the
406 LS and SL conditions in the original study was $r = .52$, we set the expected correlation
407 between repeated measurements to $\rho = .50$ in our power analysis.

408 This setup results in a required sample size of 92 participants for the complete 2 x 2
409 design. From this number, the sample size of the original study ($n = 23$) can be subtracted,
410 since these data are already available and can be included in the ANOVA model. Hence, the
411 required sample size for our replication study was only $92 - 23 = 69$ participants. We chose
412 to oversample slightly to account for the possibility that some participants need to be
413 excluded from the analysis, so the target sample size for our replication study was
414 preregistered as $N = 80$. All participants were recruited via the online platform Prolific
415 (prolific.com). Only native English speakers with a minimum approval rate of 95% and a
416 maximum number of 30 previous submissions on the Prolific platform were allowed to
417 participate. Additionally, they were required to be currently studying according to Prolific's
418 prescreening. Participants were paid 8.13£ for a study duration of about 65 minutes.

419 Our preregistered exclusion criteria included (a) obtaining a mean free recall accuracy

420 of 0 across both conditions, or (b) answering “no” to a seriousness check at the end of the
421 study (“Please tell us whether you have taken part seriously, so that we can use your answers
422 for our scientific analysis”).

423 $N = 80$ individuals participated in the experiment. According to our exclusion
424 criteria, no participants had to be excluded from the final sample. Mean age was 22.28 years
425 ($SD = 4.16$, $range = 18-42$). 62 participants (77.50%) indicated to be female, 18 participants
426 (22.50%) indicated to be male. 15 participants (18.75%) were enrolled in a
427 psychology-related study program at the time of participation.

428 *Design*

429 We used a simple study design with two L2 timing conditions: LS (240-sec L1-L2
430 interval, 60-sec L2-T interval) and SL (60-sec L1-L2 interval, 240-sec L2-LT interval). L2
431 timing was manipulated within-participants with four trials per condition, that is, eight
432 trials in total.

433 *Material*

434 For creating the word pool used to generate L1 and L2 learning lists, we exactly
435 followed the approach by Ecker, Brown, et al. (2015). 320 words were taken from the
436 Medical Research Council (MRC) Psycholinguistic Database (Wilson, 1988). All of them
437 satisfy the following criteria: They (a) are one-syllable English nouns, (b) are between 3 and
438 6 letters long, (c) have a Kucera-Francis frequency of greater than 28, and (d) have
439 familiarity and concreteness ratings of at least 400. From this word pool, 16 lists containing
440 10 words each were generated randomly for each participant, with the restriction that for
441 both lists of the same trial, no adjacent words from the alphabetically ordered word pool
442 were chosen. In addition, two word lists were created for a practice trial. These words had
443 the same characteristics as those for the experimental trials, except for a lower
444 Kucera-Francis frequency of 25-28.

445 *Procedure*

446 In contrast to the original study by Ecker, Brown, et al. (2015), our replication study
447 was conducted within one single session. Each participant completed eight trials, that is,
448 four trials per condition. The order of LS and SL trials was randomly determined for each
449 participant. Each trial consisted of an original learning phase (L1), an interpolated learning
450 phase (L2), and a test phase (T). In both L1 and L2 phases, 10 words were presented
451 centrally on the computer screen for 2000 ms each, separated by an inter-stimulus interval
452 (ISI) of 400 ms. Given the lower experimental control that comes with an online setting, we
453 chose to use a longer presentation time than the 1000 ms in the original study to avoid floor
454 effects in recall accuracies (see the pilot study described in the preregistration). In the test
455 phase, participants were given a maximum of 40 seconds to type in all L1 words they could
456 remember. In the original study, recall was conducted verbally through a microphone, an
457 approach we deemed impractical for an online setting. However, to adhere as closely as
458 possible to the original procedure, participants were required to type in and confirm each
459 word individually such that previous entries were never visible on the screen. After 20
460 seconds had passed, a button appeared on the screen that allowed participants to end this
461 part of the test phase early in case they could not remember any more L1 words. In half of
462 all trials, L1 free recall was followed by an L2 free recall test. These trials were randomly
463 selected for each participant with the restriction that no more than two successive trials
464 could omit the L2 test. Again, participants had 40 seconds and could end the test early after
465 20 seconds.

466 During the L1-L2 and L2-T intervals, participants engaged in a color-detection task
467 that was designed to mimic the tone-detection task used by Ecker, Brown, et al. (2015). We
468 opted against the original distractor task to avoid technical issues relating to audio settings
469 of participants' browsers. Colored squares were presented centrally on the computer screen,
470 each for a duration of 150 ms and with an ISI of 600 ms. Squares could be either grey (about

471 80% probability) or blue (about 20% probability). Participants were asked to press the “H”
472 key on their keyboard as quickly as possible whenever a blue square appeared on the screen.
473 Whenever participants failed to react to a target square, the text “MISS” appeared in red
474 letters on the screen for 150 ms. Whenever participants erroneously pressed the “H” key
475 when presented with a distractor square, the text “FALSE ALARM” appeared on the screen,
476 also in red letters and for 150 ms. In “long” intervals, the distractor task lasted about 240
477 seconds, whereas in “short” intervals, it only lasted 60 seconds.

478 Before the first experimental trial, participants provided demographic information
479 and were asked to answer simple questions of understanding concerning the instructions for
480 both the learning and the color-detection task. Afterwards, they completed a short practice
481 trial that contained all elements of the experimental trials. After the last trial, participants
482 were asked in a post-experimental questionnaire whether they worked on the study seriously,
483 whether they experienced any distractions or technical difficulties, and whether they used
484 any external aids (e.g., taking notes or pictures).

485 During the study, participants were immediately excluded if they (a) changed their
486 browser window more than twice, (b) stayed inactive for a considerable amount of time
487 (between 60 and 120 seconds, depending on the current part of the study, or failing to react
488 to five target squares in a row in the color-detection task), or (c) failed to answer easy
489 questions of understanding regarding the instructions twice.

490 The experiment was built in lab.js (Henninger et al., 2022) and hosted through
491 JATOS (Lange et al., 2015).

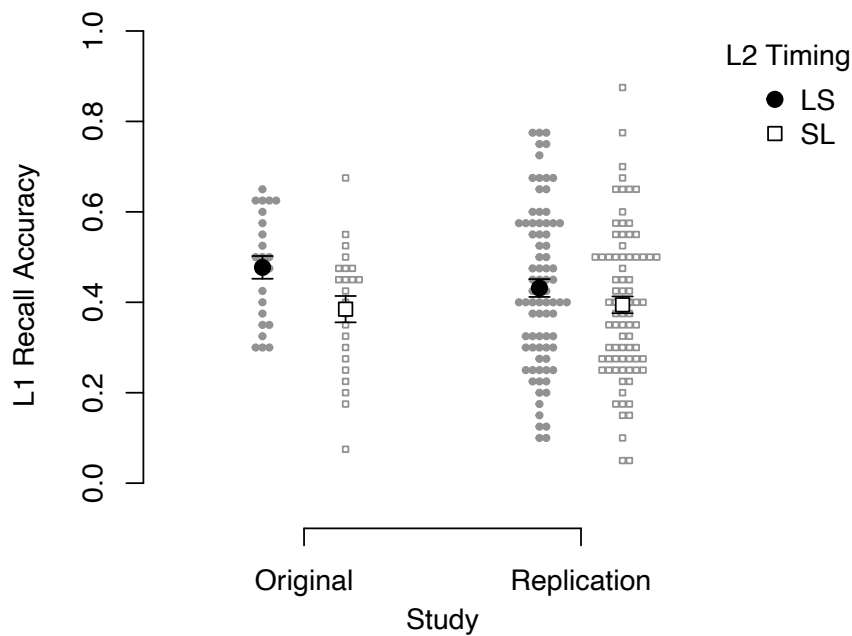
492 **Results**

493 For each participant, separate LS and SL recall accuracy scores were calculated by
494 averaging the share of correct answers in the L1 free recall test across all four trials of the
495 respective condition. Descriptively, participants in our replication study performed better in

Figure 3

L1 Recall Accuracies in Ecker, Brown, et al. (2015, Experiment 1) and Current Experiment

1



Note. Original data from Ecker, Brown, et al. (2015, Experiment 1) on the left, current replication data on the right. Original data was provided by Ullrich Ecker via personal correspondence. L2 timing was manipulated such that "long" intervals had a duration of 240 seconds, whereas "short" intervals had a duration of 60 seconds. Small points and squares represent participant-specific recall accuracies across all trials of the respective condition, large points and squares represent means, error bars represent standard errors of the mean.

496 the LS than in the SL condition, $M_{LS} = 0.43$ ($SD = 0.18$), $M_{SL} = 0.39$ ($SD = 0.17$), with a
497 small effect size of Hedges' $\hat{g} = 0.22$.

498 To determine whether this descriptive result is consistent with the Ecker, Brown, et
499 al. (2015) data, we subjected our replication data to a joint analysis with the data from the
500 original study. Figure 3 illustrates L1 free recall accuracies in the joint data set. We ran a
501 2×2 mixed ANOVA with the two factors "study" (original vs. replication) and "L2 timing"
502 (LS vs. SL) to test the null hypothesis of no interaction between these factors. This analysis
503 yielded no significant between-within interaction at $\alpha = .10$, $F(1, 101) = 2.69$, $p = .104$.
504 Thus, the null hypothesis that the TGRI we found in our replication study is equal in size to
505 the effect found in the original study could not be rejected. Note that, although irrelevant to
506 our preregistered criterion of replication success, a paired t -test for a difference between
507 conditions within our replication data turned out to be significant, $t(79) = 2.28$, $p = .013$.

508 To scrutinize the robustness of this finding, we ran a sensitivity analysis with respect
509 to distractions, technical difficulties, and the use of external aids. In the post-experimental
510 questionnaire, 22 participants (27.50%) indicated to have been distracted during the study, 3
511 participants (3.75%) indicated to have experienced technical problems, and 2 participants
512 (2.50%) indicated to have used external aids. This sensitivity analysis revealed that the
513 result pattern is not affected from removing these participants' data from the analysis, M_{LS}
514 $= 0.42$ ($SD = 0.18$), $M_{SL} = 0.38$ ($SD = 0.18$), Hedges' $\hat{g} = 0.25$, $F(1, 77) = 1.66$, $p = .202$.

515 For exploratory reasons, we also analyzed L2 free recall accuracies by calculating LS
516 and SL accuracy scores for each participant. Participants performed significantly better in
517 the LS than in the SL condition, $M_{LS} = 0.44$ ($SD = 0.18$), $M_{SL} = 0.36$ ($SD = 0.20$), Hedges'
518 $\hat{g} = 0.44$, $t(78) = 4.40$, $p < .001$.

519 **Discussion**

520 Using an online setting, we successfully replicated the effect of L2 timing on L1 free
521 recall accuracy found by Ecker, Brown, et al. (2015). This result suggests that the paradigm
522 proposed by these authors represents a suitable basis for further investigations of the TGRI.
523 Moreover, the successful switch from the lab to an online setting, including necessary
524 changes to the material and procedure, implies a rather high robustness of the effect within
525 this paradigm, given sufficiently high statistical power. The result from our sensitivity
526 analysis corroborates this conclusion.

527 These results nicely align with those by Landauer (1974) in the sense that waking rest
528 during L1-L2 and L2-T intervals does not seem to be a necessary precondition for the TGRI
529 to emerge. Instead, a relative difference in encoding demands between L1 and L2 learning on
530 the one hand and the distractor task on the other hand seems to be sufficient. Indeed, it has
531 been argued that the repeated encoding of the same distractor stimuli (i.e., grey and blue
532 squares in our color-detection task) is unlikely to cause forgetting (Ecker, Tay, et al., 2015).

533 Although the results from this first experiment cannot be used to differentiate
534 between opportunistic consolidation and temporal distinctiveness explanations of the TGRI,
535 our exploratory analysis might at least be interpreted as tentative evidence in favor of
536 temporal distinctiveness theory. More specifically, the observation of better L2 recall
537 performance in the LS compared to the SL condition (i.e., for pre-study L1-L2 interval
538 durations of 240 seconds compared to 60 seconds) is in line with the finding by Ecker, Tay, et
539 al. (2015) of a pre-study rest benefit due to increased temporal isolation. In contrast, from
540 an opportunistic consolidation perspective, the L1-L2 interval duration should have no effect
541 on later L2 recall. However, L1-L2 interval duration is confounded with L2-T interval
542 duration in our study design, so any findings relating to L2 free recall accuracy should be
543 treated with caution.

Experiment 2

After successfully replicating the TGRI in Experiment 1, we aimed at disentangling storage and retrieval contributions to L1 free recall accuracy in Experiment 2. For this purpose, we adapted the Riefer and Batchelder (1995) storage-retrieval MPT model. This approach required some careful deviations from the procedure used in Experiment 1. Most importantly, we included an L1 recognition test at the end of each trial to generate a data structure that would allow for an application of our MPT model. Thus, the results of this second experiment would also be informative with respect to the robustness of the TGRI across different types of memory tests.

Based on our results from Experiment 1, we hypothesized to observe significantly better L1 free recall (Hypothesis 1) and recognition accuracies (Hypothesis 2) in the LS compared to the SL condition. Moreover, based on opportunistic consolidation theory, we expected the MPT storage probability (parameter s) to be significantly higher in the LS than the SL condition (Hypothesis 3). In contrast, based on temporal distinctiveness theory, we expected the MPT recall retrieval probability (parameter r_1) to be significantly higher in the LS than the SL condition (Hypothesis 4). Thus, the model-based results of this experiment would allow us to precisely differentiate between both theoretical accounts.

We again uploaded a detailed study protocol including all hypotheses and our analysis plan to the OSF prior to data collection (osf.io/yb7sx).

Method

Participants

We used G*Power (Faul et al., 2009) to conduct an a priori power analysis. As detailed in the preregistration and in line with the relatively small TGRI effect size observed in Experiment 1, we aimed at detecting an effect of $d = 0.2$ for L1 free recall (Hypothesis 1) and recognition (Hypothesis 2) accuracy differences between the LS and SL conditions in a

569 one-tailed repeated-measures t -test with a power of $1 - \beta = 80\%$ and $\alpha = 5\%$. This setup
570 led to a required sample size of 156 participants. To account for the possibility that some
571 participants need to be excluded from the analysis, our preregistered target sample size was
572 $N = 180$ participants.

573 All participants were recruited via Prolific (prolific.com). Only native English
574 speakers with a minimum approval rate of 95% on the Prolific platform were allowed to
575 participate. Additionally, they were required to be currently enrolled in a study program.
576 Participants were paid 6.00£ for a study duration of about 45 minutes.

577 Our preregistered exclusion criteria were the same as in Experiment 1, that is, they
578 included (a) obtaining a mean free recall accuracy of 0 across both conditions, or (b)
579 answering “no” to a seriousness check at the end of the study.

580 A total of 180 individuals participated in Experiment 2. Of these, 3 participants were
581 excluded from the analysis because they either did not provide any correct responses in any
582 of the free recall tests or indicated not to have taken part seriously. Thus, the final sample
583 size was $N = 177$. Mean age was 28.64 years ($SD = 8.19$, $range = 18-70$). 90 participants
584 (50.85%) indicated to be female, 86 participants (48.59%) indicated to be male. One
585 participant did not wish to provide their gender. 13 participants (7.34%) were enrolled in a
586 psychology-related study program at the time of participation.

587 *Design*

588 The experimental design was the same as in Experiment 1, that is, there was an LS
589 (240-sec L1-L2 interval, 60-sec L2-T interval) and an SL condition (60-sec L1-L2 interval,
590 240-sec L2-T interval).

591 Experiment 2 entailed only four instead of eight trials, that is, two instead of four
592 trials per condition. We believe the loss in reliability to be justified by the necessity to keep

593 the total study duration within a realistic range despite some necessary procedural changes
594 (see Procedure section).

595 *Material*

596 The word pool used to generate L1 and L2 learning lists was the same as in
597 Experiment 1. For each trial, an L1 list, an L2 list, and a distractor list for the recognition
598 test (see Procedure section) was generated, each comprising 10 words.

599 *Procedure*

600 The procedure was very similar to the one used in Experiment 1, with one major
601 exception. Most importantly, an old-new recognition test was included at the end of each
602 trial to allow for an application of the storage-retrieval MPT model. Thus, after finishing the
603 free recall test, participants were presented with all L1 words of the respective trial, randomly
604 intermixed with the same number of new words that were not presented in any other trial.
605 For each word, participants were asked to press the “S” key on their keyboard when they
606 thought the word was “old” and the “L” key when they thought the word was “new”. In case
607 participants did not know whether a word was old or new, they were asked to guess.

608 In addition, an L2 free recall test was included in each trial (instead of only 50% of
609 trials) to increase the perceived relevance of L2 learning for participants. Thus, the test
610 phase in each trial included L1 free recall, L1 recognition, and L2 free recall.

611 During the study, participants were immediately excluded if they (a) stayed inactive
612 for a considerable amount of time (failing to react to five target squares in a row in the
613 color-detection task), or (b) failed to answer easy questions of understanding regarding the
614 instructions twice.

615 The experiment was again built in lab.js (Henninger et al., 2022) and hosted through
616 JATOS (Lange et al., 2015).

617 Results

618 *Memory Performance Measures*

619 We hypothesized that participants would show significantly better L1 free recall
620 (Hypothesis 1) and recognition accuracies (Hypothesis 2) in the LS compared to the SL
621 condition. In line with Hypothesis 1, mean L1 free recall accuracy was higher in the LS than
622 in the SL condition, $M_{LS} = 0.39$ ($SD = 0.19$), $M_{SL} = 0.37$ ($SD = 0.20$), $t(176) = 2.10$, $p =$
623 $.018$, Hedges' $\hat{g} = 0.13$. Recall accuracies are illustrated in Figure 4.

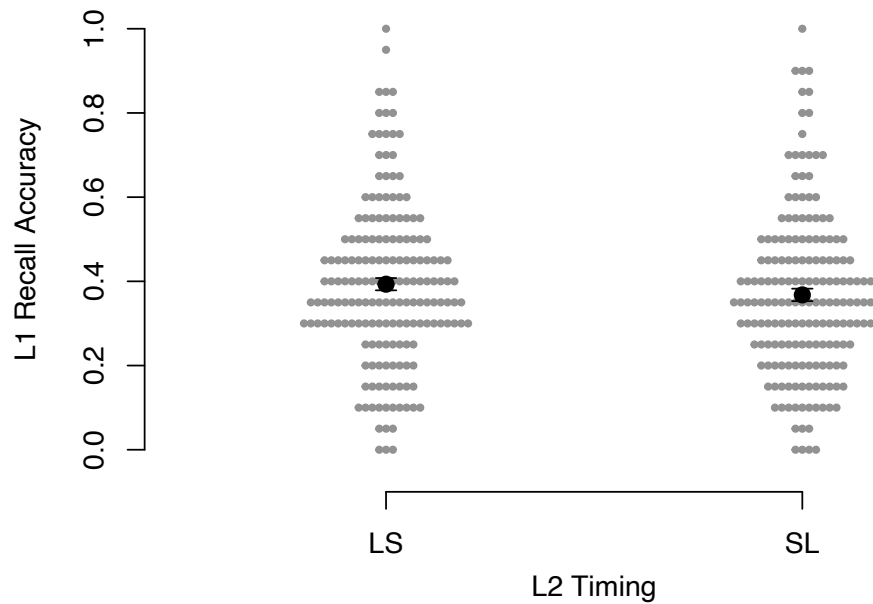
624 We calculated response-bias-corrected L1 recognition accuracy scores by subtracting
625 false alarm from hit rates for each participant (see Nadarevic, 2017). Mean recognition
626 accuracies did not differ significantly between the LS and the SL condition, $M_{LS} = 0.59$ (SD
627 $= 0.24$), $M_{SL} = 0.59$ ($SD = 0.27$), $t(176) = -0.16$, $p = .562$, Hedges' $\hat{g} = -0.01$. Thus,
628 Hypothesis 2 was not supported by the data.

629 In line with our exploratory analysis of Experiment 1, we scrutinized L2 free recall
630 accuracy in both conditions. Participants performed significantly better in the LS than in
631 the SL condition, $M_{LS} = 0.39$ ($SD = 0.22$), $M_{SL} = 0.33$ ($SD = 0.25$), Hedges' $\hat{g} = 0.29$,
632 $t(176) = 4.40$, $p < .001$.

633 *Model-Based Results*

634 Based on opportunistic consolidation theory, we expected MPT storage probabilities
635 (i.e., parameter s) to be significantly higher in the LS compared to the SL condition
636 (Hypothesis 3). Likewise, based on temporal distinctiveness theory, we expected MPT recall
637 retrieval probabilities (i.e., parameter r_1) to be significantly higher in the LS compared to
638 the SL condition (Hypothesis 4).

639 We fitted the MPT model using both aggregated and individual category frequencies
640 to confirm the robustness of our model-based results. For the aggregated data, we used a

Figure 4*L1 Recall Accuracies in Experiment 2*

Note. Participants were presented with 10 words per trial. Timing of the interpolated learning phase (L2) was manipulated such that "long" intervals had a duration of 240 seconds, whereas "short" intervals had a duration of 60 seconds. Small points represent participant-specific recall accuracies across all trials of the respective condition, large points represent means, error bars represent standard errors of the mean.

641 maximum likelihood approach as implemented in the R package MPTinR (Singmann &
642 Kellen, 2013), whereas for the individual data, we used a Bayesian hierarchical latent-trait
643 approach (Klauer, 2010) as implemented in the R package TreeBUGS (Heck et al., 2018).
644 Whereas the first estimation approach assumes observations to be identically and
645 independently distributed (i.i.d.), the latter accounts for the potential heterogeneity of
646 participants and explicitly includes parameter correlations (Heck et al., 2018).

647 For both estimation approaches, we restricted the guessing parameter g to be equal
648 between conditions (i.e., $g_{LS} = g_{SL}$) and set the distractor detection parameter $d = s * r_2$
649 (see Nadarevic, 2017) to obtain an identifiable model version. This preregistered model
650 specification yielded a good fit to the aggregated data, $G^2(1) = 1.47$, $p = .225$. To further
651 simplify the model, we also tried to fit an even more parsimonious model version with $d = s$.
652 Since this model fit the aggregated data just as well, $G^2(1) = 0.94$, $p = .331$, we used it as
653 our baseline model for further analyses. For the individual data, convergence of the MCMC
654 sampler was satisfactory as indicated by $\hat{R} < 1.05$ for all parameters (see Gelman & Rubin,
655 1992), and the model fit the data well according to posterior predictive p values, $p_1 = .490$,
656 $p_2 = .165$. Parameter estimates as well as inferences concerning parameter differences
657 between conditions nicely converged between estimation approaches. Estimates from both
658 estimation approaches are provided in Table 1. Corresponding estimates from the
659 preregistered model version with $d = s * r_2$ are provided in Table A1 in Appendix A.

660 For the aggregated data, equality constraints were imposed on parameters of interest
661 to infer the statistical significance of differences between conditions according to the test
662 statistic ΔG^2 (see Schmidt et al., 2023). For one-sided tests with $df = 1$, $z = \sqrt{\Delta G^2(1)}$ is
663 reported. For the individual data, the reliability of parameter differences corresponding to
664 one-sided hypotheses was evaluated based on the so-called Bayesian p value, that is, the
665 proportion of the posterior distribution of the respective difference estimate below zero.

666 Storage probabilities s were not significantly higher in the LS condition,

Table 1*Results of the Storage-Retrieval MPT Analysis of Experiment 2*

Parameter	LS	SL
Aggregated data ^a		
s	.61 [.59, .63]	.62 [.60, .64]
r_1	.64 [.61, .67]	.60 [.57, .62]
r_2	.93 [.91, .95]	.92 [.90, .94]
g	.41 [.40, .43]	.41 [.40, .43]
Individual data ^b		
s	.63 [.59, .67]	.65 [.61, .69]
r_1	.66 [.62, .71]	.60 [.56, .65]
r_2	.97 [.95, .99]	.95 [.92, .98]
g	.36 [.32, .40]	.36 [.32, .40]

Note. LS = "Long-Short" (i.e., long L1-L2 interval, short L2-T interval), SL = "Short-Long" (i.e., short L1-L2 interval, long L2-T interval). Parameter s = probability of storing an L1 word, r_1 = probability of retrieving an L1 word during recall, r_2 = probability of retrieving an L1 word during recognition, g = probability of guessing "old" during recognition.

Parameter g was restricted to be equal between conditions.

^a The model was fitted to the aggregated category frequencies using maximum likelihood (ML) estimation in the R package MPTinR (Singmann & Kellen, 2013). 95% confidence intervals are indicated in brackets.

^b The model was fitted to the individual category frequencies using Bayesian hierarchical estimation in the R package TreeBUGS (Heck et al., 2018). 95% Bayesian credibility intervals are indicated in brackets.

667 $z = \sqrt{\Delta G^2(1)} = 0.47$, $p = .683$, Bayesian $p = .800$. Thus, Hypothesis 3 was not supported
 668 by the data. In contrast, recall retrieval probability r_1 was significantly higher in the LS
 669 condition compared to the SL condition, $z = \sqrt{\Delta G^2(1)} = 2.25$, $p = .012$, Bayesian $p = .012$,
 670 confirming Hypothesis 4. Recognition retrieval probabilities r_2 did not differ significantly,
 671 $z = \sqrt{\Delta G^2(1)} = 0.87$, $p = .191$, Bayesian $p = .130$.

672 *Sensitivity Analysis*

673 We ran a sensitivity analysis to test the robustness of our findings. In this analysis,
 674 data from participants who changed the browser window more than twice ($n = 4$, 2.26%)
 675 was excluded. Data from participants who reported in the post-experimental questionnaire
 676 to have been distracted during the study ($n = 12$, 6.78%) or to have experienced technical
 677 problems ($n = 6$, 3.39%) was excluded as well. No participant stayed inactive for a
 678 considerable amount of time at least once during the recognition test, so this preregistered
 679 sensitivity criterion did not need to be considered. The same held true for the use of external
 680 aids.

681 The sensitivity analysis revealed that the result pattern was not affected by removing
 682 these participants' data from either the design-based or the model-based analysis. More
 683 specifically, whereas L1 free recall accuracy was significantly higher in the LS compared to
 684 the SL condition, $M_{LS} = 0.39$ ($SD = 0.20$), $M_{SL} = 0.37$ ($SD = 0.20$), $t(161) = 1.96$, $p =$
 685 $.026$, Hedges' $\hat{g} = 0.12$, no significant difference emerged for L1 recognition accuracy, $M_{LS} =$
 686 0.59 ($SD = 0.25$), $M_{SL} = 0.59$ ($SD = 0.27$), $t(161) = -0.30$, $p = .618$, Hedges' $\hat{g} = -0.02$.
 687 Moreover, whereas storage probabilities s did not differ significantly between conditions, the
 688 benefit for recall retrieval probabilities r_1 in the LS condition was robust. This held true for
 689 the aggregated data as well as for the individual data. More specifically, for the aggregated
 690 data, estimates for parameter s were $.61$ (95% CI $[.59, .63]$) in the LS and $.63$ (95% CI $[.61,$
 691 $.65]$) in the SL condition, $z = \sqrt{\Delta G^2(1)} = 0.94$, $p = .827$. Estimates for parameter r_1 were
 692 $.64$ (95% CI $[.61, .67]$) in the LS and $.59$ (95% CI $[.56, .62]$) in the SL condition,

693 $z = \sqrt{\Delta G^2(1)} = 2.55$, $p = .005$. Likewise, for the individual data, estimates for parameter s
694 were .63 (95% BCI [.59, .68]) in the LS and .66 (95% BCI [.61, .71]) in the SL condition,
695 Bayesian $p = .898$. Estimates for parameter r_1 were .66 (95% BCI [.61, .72]) in the LS and
696 .60 (95% BCI [.55, .64]) in the SL condition, Bayesian $p = .007$.

697 Discussion

698 In Experiment 2, we again successfully replicated the TGRI in free recall, that is, L1
699 free recall accuracies were significantly higher in the LS compared to the SL condition. In
700 contrast, recognition accuracy was not at all affected by L2 timing. Thus, the type of
701 memory test applied seems to be a relevant moderator of the TGRI.

702 An inspection of our MPT results provides clear evidence for a recall retrieval
703 (parameter r_1) advantage in the LS compared to the SL condition, whereas storage
704 probabilities (parameter s) did not differ between conditions. Thus, the TGRI we observed
705 on a surface level in recall accuracies was the product of underlying retrieval processes,
706 without any storage contribution. This unambiguous result pattern proved to be robust in
707 our sensitivity analysis. Thus, our results from Experiment 2 provided clear-cut evidence in
708 favor of temporal distinctiveness theory. This interpretation was again complemented by
709 better L2 recall accuracies in the LS than in the SL condition.

710 The model-based results help explain the diverging pattern with respect to L1 free
711 recall and recognition accuracy differences between LS and SL conditions. Successful
712 recognition of a previously studied item can be assumed to rely almost exclusively on this
713 item being stored in memory, whereas retrieval should only play a negligible role (Schonfield
714 & Robertson, 1966). Indeed, this assumption is confirmed by recognition retrieval
715 probabilities (parameter r_2) close to 1 in both conditions of our experiment. From this
716 perspective, the absence of an L2 timing effect on recognition accuracy is a necessary
717 consequence of storage probabilities not differing between experimental conditions. In

718 contrast, free recall tests as applied in our experiment provide ideal conditions for underlying
719 recall retrieval effects to emerge.

720 Whereas the results from Experiment 2 are fully in line with a temporal
721 distinctiveness account of the TGRI, they are hard to reconcile with opportunistic
722 consolidation theory. Indeed, any difference in consolidation between conditions should have
723 become apparent in our MPT storage parameter s . That being said, proponents of
724 opportunistic consolidation theory might in principle argue that the online setting of our
725 experiment was not appropriate to investigate consolidation effects in the first place. Indeed,
726 any environment outside the lab might impose higher-than-intended encoding demands on
727 participants. Although data quality in online settings has been shown to be generally
728 comparable or even superior to that achieved in laboratory settings (see Hartshorne et al.,
729 2019), consolidation processes might be especially susceptible to even relatively minor
730 distractions. Note, however, that reports of distractions in the post-experimental
731 questionnaire were scarce ($n = 12$, 6.78% of the total sample). Nevertheless, we cannot fully
732 rule out the possibility that consolidation was inhibited during L1-L2 and L2-T intervals of
733 our online experiment. Thus, while the results from Experiment 2 provide clear evidence in
734 favor of temporal distinctiveness theory, they do not necessarily exclude the possibility of an
735 additional opportunistic consolidation effect.

736

Experiment 3

737 Our primary aim in Experiment 3 was to replicate our findings from Experiment 2 in
738 a more controlled lab environment, thereby avoiding any unintended inhibition of
739 consolidation processes during the L1-L2 and L2-T intervals. Thus, we were interested to see
740 whether a switch from an online to a lab setting would result in differences in MPT storage
741 probabilities (parameter s) between conditions.

742 One hypothetical data pattern to be expected in Experiment 3 might have been an

743 L2 timing effect on storage *and* recall retrieval probabilities. In line with our theoretical
744 reasoning, such a pattern could be easily explained by assuming simultaneous contributions
745 of opportunistic consolidation and temporal distinctiveness to memory performance.
746 However, we wanted our results from Experiment 3 to be informative with respect to an
747 alternative explanation of such a hypothetical result pattern as well: Whereas synaptic
748 consolidation should primarily be reflected in storage parameter s , the integration of new
749 memories into pre-existing memory networks through systems consolidation might
750 additionally influence recall retrieval parameter r_1 . Although we deem the relevance of
751 systems consolidation to be negligible on a time scale of a few minutes, others have argued
752 that synaptic and systems consolidation might be coupled processes that both benefit from
753 post-encoding interference minimization (see Mednick et al., 2011). From such a perspective,
754 opportunistic synaptic and systems consolidation would be sufficient to explain a
755 simultaneous L2 timing effect on storage and recall retrieval probabilities, without a need for
756 an additional distinctiveness mechanism.

757 To avoid an ambiguous result pattern, we adapted our experimental design from
758 Experiments 1 and 2 to include “L1-L2 similarity” as a second factor besides “L2 timing”.
759 Thus, we manipulated L2 items to be either similar or dissimilar to L1 items (words as L1
760 and L2 items vs. words as L1 items and geometric figures as L2 items). The resulting 2 x 2
761 design mirrors the design used by Mercer (2015). Thereby, the results from our
762 storage-retrieval MPT analyses would help explain his findings. Crucially, for our purposes,
763 the L1-L2 similarity factor allowed for a differentiation of generalized distinctiveness and
764 opportunistic systems consolidation effects on recall retrieval probabilities (parameter r_1) in
765 case of an L2 timing effect on both storage and recall retrieval probabilities. On the one
766 hand, L1-L2 similarity should be irrelevant from an opportunistic consolidation perspective,
767 since any interpolated encoding demands are assumed to inhibit consolidation processes
768 regardless of similarity (see Dewar et al., 2007). On the other hand, from a generalized
769 distinctiveness perspective (see Ecker, Brown, et al., 2015) that assumes the temporal and

770 similarity dimensions to be equally important, very low L1-L2 similarity might be expected
771 to preclude retroactive interference effects, thereby attenuating differences between LS and
772 SL conditions.

773 Against this backdrop, the following hypotheses may be derived from opportunistic
774 consolidation and distinctiveness theory: Both theoretical accounts imply significantly higher
775 L1 free recall accuracies in the LS than in the SL conditions (Hypothesis 1). From an
776 opportunistic consolidation perspective, the TGRI should generalize to L1 recognition
777 accuracies (i.e., significantly higher L1 recognition accuracies in the LS compared to the SL
778 conditions; Hypothesis 2), and both effects should be driven by significantly higher MPT
779 storage probabilities (parameter s) in the LS compared to the SL conditions (Hypothesis 3).

780 In contrast, the predictions to be derived from a generalized distinctiveness theory
781 depend on the relative importance of the temporal versus similarity dimensions of the
782 hypothesized latent memory space. If both dimensions contribute more or less equally to
783 retroactive interference effects, an interaction effect of L2 timing and L1-L2 similarity on
784 MPT recall retrieval probabilities (parameter r_1) may be observed. Specifically, the
785 difference in parameter r_1 between the LS and SL conditions should be less pronounced in
786 the low than in the high L1-L2 similarity conditions (Hypothesis 4).

787 If, however, the importance of the similarity dimension would turn out to be
788 unexpectedly low (reflected, in the most extreme case, by a non-significant main effect of
789 L1-L2 similarity on L1 free recall), a mere main effect of L2 timing on parameter r_1 would be
790 expected (Hypothesis 5). In the absence of a significant L2 timing effect on parameter s (see
791 Hypothesis 3), such an observation would be best explained by a purely *temporal*
792 distinctiveness theory.

793 The corresponding preregistration including a study protocol, all hypotheses², and
794 the analysis plan is available on the OSF (osf.io/za8yt).

795 Method

796 *Participants*

797 We conducted an a priori power analysis in multiTree (Moshagen, 2010). We aimed
798 at a high power of $1 - \beta = 95\%$ to detect an interaction effect of L2 timing and L1-L2
799 similarity on MPT parameter r_1 (see Hypothesis 4). To implement such a test, shrinkage
800 parameters α_{LS} and α_{SL} were used to reparameterize parameter r_1 in both “high L1-L2
801 similarity” conditions (see Kuhlmann et al., 2019 for details on MPT interaction tests).
802 Specifically, in the LS condition, we set r_{11} (i.e., r_1 in the high similarity LS condition) equal
803 to r_{13} (i.e., r_1 in the low similarity LS condition) multiplied with α_{LS} , that is, $r_{11} = r_{13} * \alpha_{LS}$.
804 Likewise, in the SL condition, we set r_{12} (i.e., r_1 in the high similarity SL condition) equal to
805 r_{14} (i.e., r_1 in the low similarity SL condition) multiplied with α_{SL} , that is, $r_{12} = r_{14} * \alpha_{SL}$.
806 Thereby, the interaction test boiled down to a one-sided test of the equality constraint
807 $\alpha_{LS} = \alpha_{SL}$. In our power analysis, we set $\alpha_{LS} = .85$ and $\alpha_{SL} = .75$. Thus, the critical
808 difference in shrinkage parameters to be detected by the corresponding significance test was
809 $.10$. For a one-sided test with $\alpha = .05$ (corresponding to a standard two-tailed test with $\alpha =$
810 $.10$), this setup led to a required sample size of 130 participants given that each participant
811 responded to 160 items. Our target sample size was $N = 150$ participants to account for the
812 possibility that some participants need to be excluded from the analysis. A detailed overview
813 of our power analysis setup is provided in Table B1 in Appendix B.

² Note that, in the preregistration, we derived all hypotheses from either opportunistic consolidation theory or a generalized distinctiveness theory that assumes the temporal and similarity dimensions to be equally important. The hypotheses presented here more comprehensively reflect the entire range of theoretically meaningful outcomes by acknowledging the possibility that the similarity dimension may be less relevant than expected.

814 Participants were recruited via the study participation platform of the University of
815 Mannheim and through personal communication. Participants were required to be fluent in
816 German and between 18 and 30 years old. They could choose between study credit and a
817 financial compensation of 15€ for a study duration of about 90 minutes.

818 Our preregistered exclusion criteria included (a) obtaining a mean L1 free recall
819 accuracy of 0 across both conditions, (b) obtaining L1 free recall or recognition accuracies
820 more than 3 times the median absolute distance (MAD) below or above the respective
821 median (see Leys et al., 2013), or (c) answering “no” to a seriousness check at the end of the
822 study (unless the explanatory text input was deemed irrelevant).

823 A total of 150 individuals participated in Experiment 3. Of these, 10 participants
824 were excluded from the analysis because they satisfied one of our exclusion criteria,
825 completed parts of the experiment twice due to a technical error, or took notes during L1
826 and L2 learning phases. Thus, the final sample size was $N = 140$. Mean age was 22.38 years
827 ($SD = 2.95$, $range = 18-30$). 101 participants (72.14%) identified as female, 37 participants
828 (26.43%) identified as male, 1 participant identified their gender as non-binary, and 1
829 participant refrained from providing their gender. 127 participants (90.71%) were currently
830 enrolled in a study program, and 75 participants (53.57%) were enrolled in a psychology
831 program at the time of participation.

832 *Design*

833 We used a 2 x 2 design with the two factors “L2 timing” and “L1-L2 similarity”.
834 Whereas the LS conditions were the same as in Experiments 1 and 2 (240-sec L1-L2 interval,
835 60-sec L2-T interval), the L1-L2 interval was shortened from 60 to 20 seconds in the SL
836 conditions (i.e., 20-sec L1-L2 interval, 280-sec L2-T interval) in an attempt to increase the
837 effect size of L2 timing on L1 free recall accuracy. L1-L2 similarity was manipulated such
838 that words were presented for L1 and L2 learning (high similarity), or words were presented

839 for L1 learning and geometric figures for L2 learning (low similarity). Both factors were
840 manipulated within-participants with two trials per condition, that is, each participant
841 completed eight trials in total.

842 *Material*

843 The word pool used to generate L1 and L2 learning and recognition distractor lists
844 consisted of 200 concrete German nouns taken from Hager and Hasselhorn (1994). All words
845 were 1-2 syllables and 3-6 letters long and had concreteness ratings ≥ 0.8 on a scale from -20
846 (very abstract) to 20 (very concrete). Words for the practice trial were selected from the
847 same source and had concreteness ratings between -2.43 and 0.8. For each trial, an L1 list
848 and a distractor list was generated. L2 word lists were only generated for half of the trials,
849 that is, for the high L1-L2 similarity trials. Each list consisted of 10 words.

850 For L2 lists in the low L1-L2 similarity conditions and the practice trial, we generated
851 a pool of 50 geometric figures. Our aim was to create items that would be hard to verbalize
852 for participants to minimize their potential similarity to L1 words. All figures consisted of
853 some combination of circles, squares, triangles, pentagons, straight or curved lines, and
854 straight or curved arrows of different sizes and orientations. Figures were organized into five
855 groups according to their main element (circle, square, equilateral triangle, right-angled
856 triangle, pentagon). Each L2 list consisted of 10 geometric figures and contained two
857 randomly selected figures from each group. Some example stimuli are presented in Figure 5.

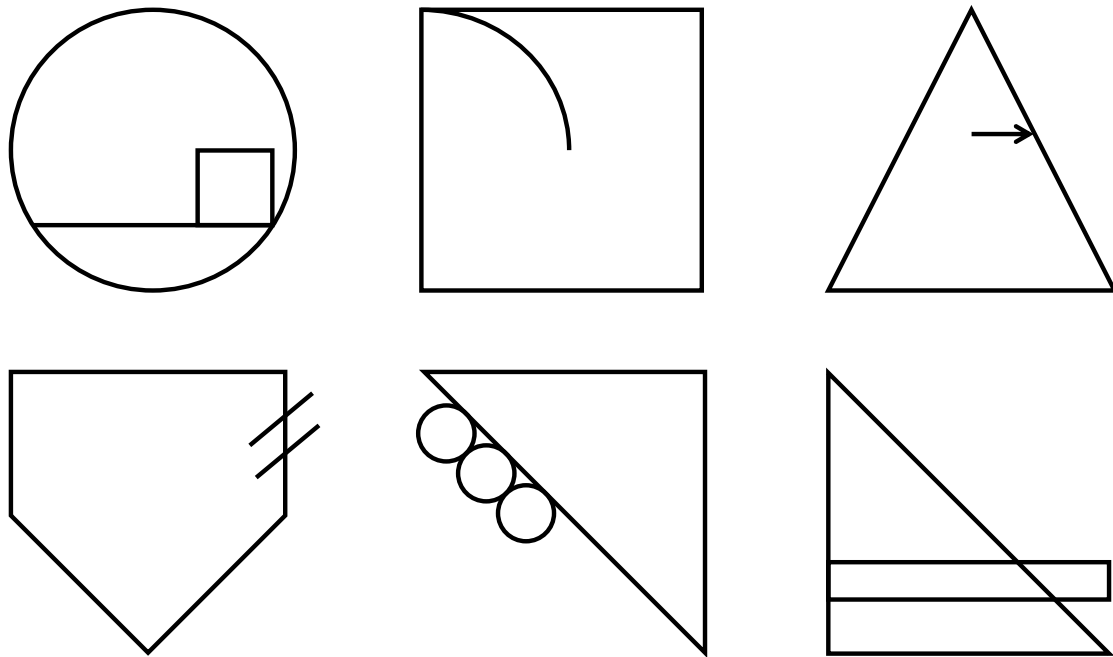
858 *Procedure*

859 The basic procedure largely followed that from Experiment 2, with some minor
860 exceptions. First, participants listened to ambient music for just under three minutes at the
861 very beginning of their participation. This introduction was intended to put participants in a
862 relaxed and focused state. The practice trial included geometric figures for L2 learning.

863 L1 items had the same presentation time as in Experiments 1 and 2, that is, 2000 ms

Figure 5

Examples of Geometric Figures Used as L2 Items in Low L1-L2 Similarity Conditions in Experiment 3



Note. Each L2 list consisted of 10 geometric figures.

864 each. In contrast, during L2 learning, items were presented for 4000 ms each to ensure that
865 geometric figures in the low L1-L2 similarity conditions could be encoded sufficiently well
866 (see the pilot studies described in the preregistration). Geometric figures were presented on a
867 white background.

868 We used an adapted version of the original tone-detection task by Ecker, Brown, et al.
869 (2015) as a distractor task during the L1-L2 and L2-T intervals. A randomly ordered
870 sequence of low (440 Hz, i.e., note A4) and high tones (523 Hz, i.e., note C5) was presented
871 to participants via headphones, each for 150 ms and with an ISI of 1000 ms. Low tones were
872 presented with a probability of about 80%, high tones with a probability of about 20%.
873 Participants were asked to press the “H” key on their keyboard as quickly as possible
874 whenever a high tone was presented. Whenever participants failed to react to a target tone
875 or erroneously pressed the “H” key when presented with a distractor tone, a very low error
876 tone (330 Hz, i.e., note E4) was played for 150 ms. We aimed at minimizing mental effort
877 and encoding demands while keeping active rehearsal of L1 and L2 items minimal. To this
878 end, we not only used a lower presentation frequency than Ecker, Brown, et al. (2015, i.e.,
879 1000 ms instead of 600 ms ISI), but also asked participants to close their eyes during the
880 task. The end of each block of the tone-detection task was indicated verbally via headphones
881 (“Please open your eyes now”).

882 Participants were given 60 seconds per L1 and L2 free recall test. For L2 free recall in
883 the low L1-L2 similarity conditions, participants used a booklet that was placed right next to
884 the computer screen. On each page of the booklet, ten empty boxes were provided.
885 Participants were asked to draw as many of the previously presented geometric figures as
886 possible, one per box. After 60 seconds had passed, participants were told to stop drawing
887 and to turn the page of the booklet. Booklets contained six pages for one practice trial and
888 four experimental trials, that is, there was one extra page. Thereby, participants were unable
889 to predict the L1-L2 similarity condition of the last trial based on whether or not an empty

890 page remained in the booklet.

891 Each participant completed eight experimental trials, that is, four trials per factor
892 level and two trials per condition. The order of the four conditions across trials was
893 randomly determined with the only restriction that no factor level was repeated more than
894 twice in a row.

895 In a post-experimental questionnaire, participants were asked to indicate whether
896 they had taken part seriously, whether they had understood all instructions, and whether
897 they had engaged in active rehearsal of L1 and/or L2 items during the tone-detection task.
898 If they indicated to have engaged in active rehearsal, they were asked to rate the rehearsal
899 frequency on a 7-point Likert scale from “not at all” to “very often”. They were also asked to
900 indicate whether they had any guess about the background of the study.

901 The experiment was built in OpenSesame (Mathôt et al., 2012).

902 **Results**

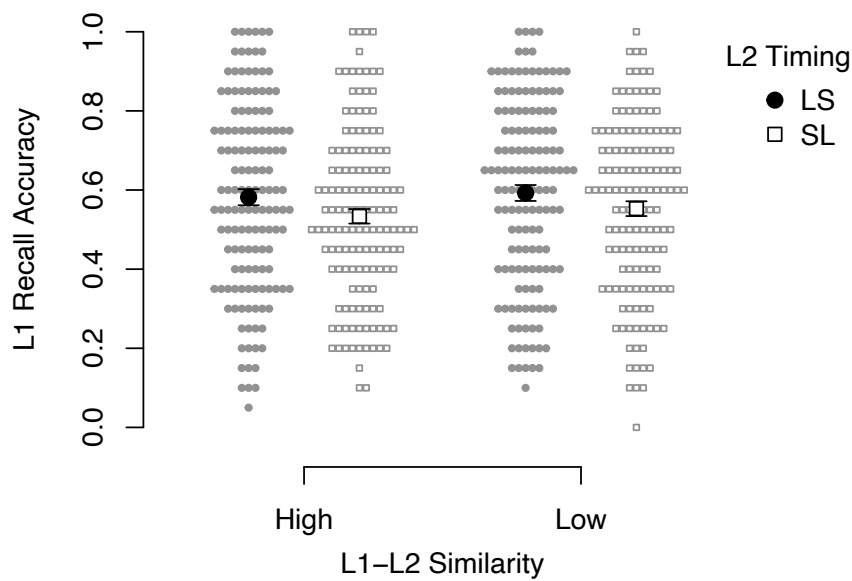
903 *Memory Performance Measures*

904 We predicted that L1 free recall accuracies would be higher in the LS compared to
905 the SL conditions (Hypothesis 1). Descriptively, this was the case both in the high L1-L2
906 similarity conditions, $M_{LS} = 0.58$ ($SD = 0.24$), $M_{SL} = 0.53$ ($SD = 0.22$), as well as in the
907 low L1-L2 similarity conditions, $M_{LS} = 0.59$ ($SD = 0.24$), $M_{SL} = 0.55$ ($SD = 0.22$). In line
908 with this, a 2 x 2 repeated-measures ANOVA yielded a significant main effect of L2 timing,
909 $F(1, 139) = 12.92$, $p < .001$, $\eta_p^2 = 0.09$. In other words, Hypothesis 1 was confirmed by the
910 data. In contrast, there was neither a significant main effect of L1-L2 similarity, $F(1, 139) =$
911 1.38 , $p = .242$, $\eta_p^2 = 0.01$, nor a significant interaction effect of both factors, $F(1, 139) =$
912 0.12 , $p = .727$, $\eta_p^2 = 0.00$. L1 free recall accuracies are illustrated in Figure 6.

913 With respect to L1 recognition accuracy, opportunistic consolidation theory predicts

Figure 6

L1 Recall Accuracies in Experiment 3



Note. Participants were presented with 10 words per trial. Similarity between original (L1) and interpolated (L2) learning material was manipulated as "high" (i.e., words as L1 and L2 materials) versus "low" (i.e., words as L1 material, geometric forms as L2 material). L2 timing was manipulated such that in the "Long-Short" (LS) condition, the L1-L2 interval lasted 240 seconds, whereas the L2-T interval lasted 60 seconds. In contrast, in the "Short-Long" (SL) condition, the L1-L2 interval lasted 20 seconds, whereas the L2-T interval lasted 280 seconds. Small points and squares represent participant-specific recall accuracies across all trials of the respective condition, large points and squares represent means, error bars represent standard errors of the mean.

914 higher accuracies in the LS compared to the SL conditions (Hypothesis 2). Descriptively,
 915 recognition accuracies were very similar both in the high L1-L2 similarity conditions, $M_{LS} =$
 916 0.83 ($SD = 0.16$), $M_{SL} = 0.83$ ($SD = 0.15$), as well as in the low L1-L2 similarity conditions,
 917 $M_{LS} = 0.84$ ($SD = 0.14$), $M_{SL} = 0.83$ ($SD = 0.14$). Indeed, a 2 x 2 repeated-measures
 918 ANOVA yielded neither a significant main effect of L2 timing, $F(1, 139) = 1.01$, $p = .317$, $\eta_p^2 =$
 919 0.01 , nor of L1-L2 similarity, $F(1, 139) = 1.08$, $p = .299$, $\eta_p^2 = 0.01$. Thus, Hypothesis 2
 920 was rejected. Note that there was no significant interaction effect either, $F(1, 139) = 0.37$, p
 921 $= .544$, $\eta_p^2 = 0.00$.

922 We analyzed the effects of L2 timing and L1-L2 similarity on L2 free recall accuracy.
 923 Descriptively, participants performed better in the LS than in the SL conditions, both for
 924 words (i.e., high L1-L2 similarity conditions), $M_{LS} = 0.63$ ($SD = 0.26$), $M_{SL} = 0.53$ ($SD =$
 925 0.26), as well as for geometric figures (i.e., low L1-L2 similarity conditions), $M_{LS} = 0.38$ (SD
 926 $= 0.18$), $M_{SL} = 0.35$ ($SD = 0.16$). A 2 x 2 repeated-measures ANOVA revealed significant
 927 main effects of L2 timing, $F(1, 139) = 38.49$, $p < .001$, $\eta_p^2 = 0.22$, and L1-L2 similarity, $F(1,$
 928 $139) = 139.76$, $p < .001$, $\eta_p^2 = 0.50$, as well as a significant interaction effect, $F(1, 139) =$
 929 8.68 , $p = .004$, $\eta_p^2 = 0.06$.

930 ***Model-Based Results***

931 As in Experiment 2, we fitted the MPT model using both aggregated category
 932 frequencies in MPTinR (Singmann & Kellen, 2013) and individual category frequencies in
 933 TreeBUGS (Heck et al., 2018). Again, guessing parameter g was set equal between all
 934 conditions. In line with our approach in Experiment 2, we equated the distractor detection
 935 parameter d with parameter s (i.e., $d = s$). In Experiment 3, this model specification yielded
 936 a good fit to the aggregated data, $G^2(3) = 1.31$, $p = .726$. For the individual data,
 937 convergence and model fit indices were satisfactory as well, all $\hat{R} < 1.05$, $p_1 = .571$, $p_2 =$
 938 $.619$. Estimates for all parameters from both estimation approaches are provided in Table 2.

Table 2*Results of the Storage-Retrieval MPT Analysis of Experiment 3*

Parameter	High similarity		Low similarity	
	LS	SL	LS	SL
Aggregated data ^a				
<i>s</i>	.84 [.82, .85]	.83 [.82, .85]	.85 [.83, .86]	.84 [.82, .85]
<i>r</i> ₁	.70 [.68, .72]	.64 [.62, .66]	.70 [.68, .72]	.66 [.64, .68]
<i>r</i> ₂	.99 [.98, .99]	.99 [.98, .99]	.99 [.99, .99]	.99 [.99, .99]
<i>g</i>	.37 [.35, .39]	.37 [.35, .39]	.37 [.35, .39]	.37 [.35, .39]
Individual data ^b				
<i>s</i>	.88 [.85, .90]	.86 [.84, .89]	.88 [.85, .90]	.86 [.84, .89]
<i>r</i> ₁	.71 [.66, .75]	.65 [.61, .70]	.72 [.67, .76]	.67 [.63, .71]
<i>r</i> ₂	.99 [.98, .99]	.99 [.98, .99]	.99 [.99, .99]	.99 [.99, .99]
<i>g</i>	.33 [.28, .38]	.33 [.28, .38]	.33 [.28, .38]	.33 [.28, .38]

Note. LS = "Long-Short" (i.e., long L1-L2 interval, short L2-T interval), SL = "Short-Long" (i.e., short L1-L2 interval, long L2-T interval). Parameter *s* = probability of storing an L1 word, *r*₁ = probability of retrieving an L1 word during recall, *r*₂ = probability of retrieving an L1 word during recognition, *g* = probability of guessing 'old' during recognition.

Parameter *g* was restricted to be equal between conditions.

^a The model was fitted to the aggregated category frequencies using maximum likelihood (ML) estimation in the R package MPTinR (Singmann & Kellen, 2013). 95% confidence intervals are indicated in brackets.

^b The model was fitted to the individual category frequencies using Bayesian hierarchical estimation in the R package TreeBUGS (Heck et al., 2018). 95% Bayesian credibility intervals are indicated in brackets.

939 According to opportunistic consolidation theory, storage probabilities s should be
940 significantly higher in the LS than in the SL conditions (Hypothesis 3). Based on different
941 variants of distinctiveness theory, either an interaction effect (Hypothesis 4) or an L2 timing
942 main effect on recall retrieval probabilities r_1 should emerge (Hypothesis 5). Parameter
943 estimates and inferences concerning parameter differences again converged between
944 estimation approaches. Overall, for the aggregated data, storage probabilities s did not differ
945 significantly between conditions, $\Delta G^2(3) = 2.70$, $p = .440$. Thus, Hypothesis 3 was not
946 supported by the data. In contrast, recall retrieval probabilities r_1 differed significantly
947 between conditions, $\Delta G^2(3) = 19.82$, $p < .001$. A closer inspection revealed a significant
948 main effect of L2 timing on parameter r_1 , $z = \sqrt{\Delta G^2(1)} = 4.21$, $p < .001$, Bayesian $p = .014$,
949 but no main effect of L1-L2 similarity, $z = \sqrt{\Delta G^2(1)} = 1.02$, $p = .155$, Bayesian $p = .333$.
950 To test a potential interaction effect of L2 timing and L1-L2 similarity on parameter r_1 , we
951 specified an equivalent model version including shrinkage parameters α_{LS} and α_{SL} (see
952 Participants section). There was no significant interaction effect, that is, no significant
953 difference in shrinkage parameters α_{LS} and α_{SL} , $z = \sqrt{\Delta G^2(1)} = 1.09$, $p = .139$, Bayesian p
954 $= .563$. In other words, whereas Hypothesis 4 had to be rejected, Hypothesis 5 was
955 confirmed by the data. Recognition retrieval probabilities r_2 did not differ significantly
956 between conditions, $\Delta G^2(3) = 3.08$, $p = .380$.

957 *Sensitivity Analysis*

958 We conducted a sensitivity analysis to scrutinize the robustness of our results. In the
959 post-experimental questionnaire, $n = 8$ participants (5.71%) indicated not to have
960 understood all instructions, $n = 4$ participants (2.86%) had correct assumptions about the
961 background of the study, and $n = 119$ participants (85%) indicated to have engaged in active
962 rehearsal of L1 or L2 items during the tone-detection task. Given the higher-than-expected
963 occurrence of active rehearsal, we decided against our preregistered plan of excluding all
964 respective participants from the sensitivity analysis. Instead, to have a sufficiently large

965 sample that would still allow for meaningful interpretations, we only excluded those
 966 participants who selected one of the two highest categories on a 7-point Likert scale of
 967 rehearsal frequency. This more liberal criterion applied to $n = 43$ participants (30.71%).

968 For L1 free recall and recognition accuracy, this sensitivity analysis revealed a result
 969 pattern in line with the main analysis. More specifically, L1 free recall accuracies were higher
 970 in the LS than in the SL conditions both in the high L1-L2 similarity conditions, $M_{LS} = 0.54$
 971 ($SD = 0.26$), $M_{SL} = 0.50$ ($SD = 0.22$), and in the low L1-L2 similarity conditions, $M_{LS} =$
 972 0.54 ($SD = 0.23$), $M_{SL} = 0.53$ ($SD = 0.23$). In line with the main analysis, the main effect of
 973 L2 timing was significant in a 2 x 2 repeated-measures ANOVA, $F(1, 101) = 4.97$, $p = .028$,
 974 $\eta_p^2 = 0.05$. In contrast, neither the main effect of L1-L2 similarity, $F(1, 101) = 1.26$, $p = .265$,
 975 $\eta_p^2 = 0.01$, nor the interaction effect, $F(1, 101) = 1.41$, $p = .237$, $\eta_p^2 = 0.01$, were statistically
 976 significant. L1 recognition accuracies were very similar both in the high L1-L2 similarity
 977 conditions, $M_{LS} = 0.77$ ($SD = 0.28$), $M_{SL} = 0.78$ ($SD = 0.26$), as well as in the low L1-L2
 978 similarity conditions, $M_{LS} = 0.79$ ($SD = 0.27$), $M_{SL} = 0.78$ ($SD = 0.26$). As in the main
 979 analysis, a 2 x 2 repeated-measures ANOVA yielded neither a significant main effect of L2
 980 timing, $F(1, 101) = 0.39$, $p = .534$, $\eta_p^2 = 0.00$, nor of L1-L2 similarity, $F(1, 101) = 1.50$, $p =$
 981 $.223$, $\eta_p^2 = 0.01$, nor a significant interaction effect, $F(1, 101) = 0.29$, $p = .593$, $\eta_p^2 = 0.00$.

982 With respect to the MPT analyses, results again converged for the aggregated and
 983 individual data. Parameter estimates are provided in Table C1 in Appendix C. Storage
 984 probabilities s did not differ significantly between conditions, $\Delta G^2(3) = 1.60$, $p = .658$. In
 985 contrast, recall retrieval probabilities r_1 did differ, $\Delta G^2(3) = 10.71$, $p = .013$. Whereas there
 986 was a main effect of L2 timing, $z = \sqrt{\Delta G^2(1)} = 2.31$, $p = .010$, Bayesian $p = .018$, the main
 987 effect of L1-L2 similarity was not significant, $z = \sqrt{\Delta G^2(1)} = 0.18$, $p = .430$, Bayesian $p =$
 988 $.779$. The interaction effect of both factors did reach statistical significance for the
 989 aggregated data, $z = \sqrt{\Delta G^2(1)} = 2.31$, $p = .010$, but not for the individual data, Bayesian p
 990 $= .158$. Recognition retrieval probabilities r_2 did not differ significantly between conditions,

991 $\Delta G^2(3) = 6.17, p = .104.$

992 Given the unexpected ubiquity of active rehearsal, we conducted an exploratory
993 analysis of its influence on L1 free recall accuracy. To this end, we included participants'
994 self-reported rehearsal rating in an ANCOVA model including L2 timing and L1-L2
995 similarity as categorical factors. Those participants that had indicated not to have engaged
996 in any rehearsal were given a frequency rating of zero, all other participants had rated their
997 frequency on a 7-point Likert from 0 = "not at all" to 6 = "very often". Note that none of
998 those participants that indicated to have engaged in any rehearsal subsequently selected the
999 "not at all" category. This analysis yielded a significant main effect of rehearsal on L1 free
1000 recall accuracy, $F(1, 138) = 5.94, p = .016, \eta_p^2 = 0.04$. Most importantly, however, the main
1001 effect of L2 timing continued to be significant even when controlling for rehearsal, $F(1, 138)$
1002 $= 13.10, p < .001, \eta_p^2 = 0.09$. Also, there was no significant interaction effect of L2 timing
1003 and rehearsal, $F(1, 138) = 2.97, p = .087, \eta_p^2 = 0.02$.

1004 Discussion

1005 Our results from Experiment 3 mirror those obtained in Experiment 2: We again
1006 found a main effect of L2 timing on L1 free recall accuracies and MPT recall retrieval
1007 probabilities (parameter r_1), but not on L1 recognition accuracies or MPT storage
1008 probabilities (parameter s). In other words, the TGRI in free recall was once again purely
1009 retrieval-driven, without any storage contribution. This was the case despite our switch to a
1010 more controlled laboratory setting and a tone-detection task with a lower presentation
1011 frequency than in our previous color-detection task. Thus, Experiment 3 again supports
1012 temporal distinctiveness theory and provides convincing evidence against a role for
1013 opportunistic consolidation in the TGRI.

1014 Interestingly, the absence of any L1-L2 similarity effects on either L1 free recall, L1
1015 recognition, or any MPT parameters is also hard to reconcile with a *generalized*

1016 distinctiveness perspective on which several of our preregistered hypotheses were based. If
1017 the latent memory space proposed by Brown et al. (2007) is made up of more than just a
1018 temporal dimension, a similarity dimension seems to be a natural candidate, given the
1019 widely accepted role of L1-L2 similarity for retroactive interference (see Dewar et al., 2007
1020 for a critical discussion). Yet, while our results are perfectly in line with temporal
1021 distinctiveness theory, they do not provide any evidence for a role of L1-L2 similarity in the
1022 TGRI. Notably, our results nicely align with those by Mercer (2015) who used the same
1023 experimental design combined with paired associates instead of singletons as learning
1024 material, waking rest instead of a color-detection task during L1-L2 and L2-T intervals, and
1025 other methodological deviations from our approach. Together, his and our results suggest
1026 that the timing of an interpolated learning phase might be more predictive of retroactive
1027 interference effects than its similarity to the original learning phase, and that the exact
1028 conditions under which L1-L2 similarity might influence the TGRI should be investigated
1029 more explicitly in future research. The overall result pattern from our confirmatory main
1030 analysis was robust in our sensitivity analysis.

1031 Our exploratory analysis of L2 free recall accuracies revealed that high L1-L2
1032 similarity items (i.e., words) were remembered better than low L1-L2 similarity items (i.e.,
1033 geometric figures). This main effect of L1-L2 similarity on L2 free recall makes the absence
1034 of such an effect on L1 free recall even more surprising, since worse encoding of interpolated
1035 materials should generally be expected to result in reduced retroactive interference effects
1036 (Delprato, 2005). At the same time, geometric figures were apparently encoded sufficiently
1037 well to induce a TGRI.

1038 Finally, to our surprise, self-reported active rehearsal of L1 and L2 items during
1039 L1-L2 and L2-T intervals was rather frequent among our participants. Thus, our version of
1040 the tone-detection task from Ecker, Brown, et al. (2015) was not cognitively demanding
1041 enough to prevent participants from engaging in task-unrelated activities. In principle, the

1042 ubiquity of rehearsal might cast doubt on the interpretability of the TGRI we found in
1043 Experiment 3. However, such concerns are most likely unwarranted for the following two
1044 reasons: First, if active L1 rehearsal had been driving the L2 timing effect on L1 free recall,
1045 we would have expected to find a significant interaction effect of rehearsal and L2 timing on
1046 L1 free recall accuracy in our ANCOVA model, which was not the case. Second, more L1
1047 rehearsal in LS compared to SL conditions should have resulted in higher MPT storage
1048 probabilities (parameter s), which we did not find either. Thus, the TGRI we found in
1049 Experiment 3 was robust against L1-L2 similarity, purely retrieval-driven, and most likely
1050 not influenced by active rehearsal during L1-L2 and L2-T intervals.

1051 **General Discussion**

1052 In the current research, we aimed at scrutinizing the replicability of the TGRI in
1053 episodic memory and subjecting its proposed theoretical explanations to a severe test. To
1054 this end, we adapted Riefer and Batchelder's (1995) storage-retrieval MPT model to the
1055 TGRI paradigm proposed by Ecker, Brown, et al. (2015). Across three experiments,
1056 participants learned and retrieved word lists, with some interpolated learning either rather
1057 early or rather late during the 5-min retention interval. Whereas Experiments 1 and 2 were
1058 conducted in an online setting, Experiment 3 took place in a more controlled lab environment.
1059 Thereby, our series of experiments demonstrates the value of a close replication of a previous
1060 research finding (Experiment 1) as a starting point for methodological modifications
1061 (Experiments 2 and 3) that allow for a sophisticated cognitive modeling approach such as
1062 MPT modeling. All our experiments were publicly preregistered on the OSF.

1063 **Finding 1: The TGRI is robust if methodological precautions are considered**

1064 To evaluate the replicability of the TGRI, we used the basic paradigm introduced by
1065 Ecker, Brown, et al. (2015). Across all three experiments, we found that participants freely
1066 recalled significantly more L1 words in the LS compared to the SL condition. Corresponding
1067 standardized effect sizes were small but consistent across experiments (i.e., Hedges' $\hat{g} = 0.22$

1068 in Experiment 1, 0.13 in Experiment 2, and 0.21 in Experiment 3). Thus, the TGRI in free
1069 recall was robust against various modifications of the paradigm (e.g., online vs. lab setting,
1070 English vs. German word lists, 60-sec vs. 20-sec L1-L2 interval, color-detection
1071 vs. tone-detection distractor task). Our positive results from Experiments 1 and 2 suggest
1072 that online studies represent a viable alternative to lab studies for future TGRI research,
1073 such as has recently been reported for the waking rest effect (King & Nicosia, 2022; but also
1074 see Leetham et al., 2024).

1075 Given our replication success in the current research, the absence of a significant LS
1076 versus SL effect in Experiment 2 by Ecker, Brown, et al. (2015) discussed in the Introduction
1077 section can most likely be explained by insufficient statistical power of our reanalysis. Thus,
1078 the methodological recommendations proposed by Wixted (2004) have proven to be suitable
1079 guidelines for investigations of the TGRI: Using a procedure with reduced cognitive demands
1080 during L1-L2 and L2-T intervals and some minimal L2-T interval yields a reliable TGRI in
1081 free recall.

1082 In Experiments 2 and 3, we analyzed L1 recognition accuracy as an additional
1083 dependent variable that was not part of the original paradigm by Ecker, Brown, et al. (2015).
1084 In both experiments, recognition remained totally unaffected by our experimental
1085 manipulations. This suggests that the TGRI is moderated by the type of memory test
1086 applied, such that more retrieval-dependent memory tests (i.e., free recall) yield a TGRI, but
1087 more retrieval-independent tests (i.e., recognition) do not. Therefore, future research on the
1088 TGRI should not only adhere to the methodological recommendations by Wixted (2004), but
1089 also stick to more retrieval-dependent recall tests.

1090 **Finding 2: The TGRI is purely retrieval-driven**

1091 Given our success in replicating the TGRI in free recall, we adapted the procedure of
1092 Experiments 2 and 3 to accommodate a modification of the Riefer and Batchelder (1995)

1093 storage-retrieval MPT model. Our model-based results from both experiments suggest that
1094 the TGRI is purely retrieval-driven, that is, MPT recall retrieval probabilities were
1095 significantly higher in the LS compared to the SL condition. In contrast, MPT storage
1096 probabilities were unaffected by our L2 timing manipulation. This is in line with our
1097 observation of significant LS versus SL effects on free recall but not recognition in
1098 Experiments 2 and 3. Thus, our results provide clear-cut evidence in favor of a temporal
1099 distinctiveness explanation of the TGRI, since higher temporal isolation of L1 items should
1100 result in higher retrievability of those items. In contrast, the opportunistic consolidation
1101 theory cannot be easily reconciled with our data, because an increase in synaptic
1102 consolidation of L1 items should have strengthened the respective memory traces, resulting
1103 in higher storage probabilities and recognition accuracies.

1104 This interpretation nicely complements and extends the results by Ecker, Brown, et
1105 al. (2015) who found that incorporating a consolidation mechanism into their computational
1106 model did not improve model fit. Thereby, a role for consolidation in the TGRI has now
1107 been rejected through two independent modeling approaches. This makes it rather unlikely
1108 that either finding could just be an artifact of specific modeling choices.

1109 Proponents of an opportunistic consolidation account might argue that the specific
1110 paradigm used by Ecker, Brown, et al. (2015) and ourselves was ill-suited for a fair test of
1111 the opportunistic consolidation theory. First, whereas most waking rest studies use retention
1112 intervals of at least 8 minutes of unoccupied rest (see Wamsley, 2019 for a review), L1-L2
1113 intervals in our experiments never exceeded 4 minutes and included simple distractor tasks.
1114 Given the underspecification of the consolidation process (Ecker & Lewandowsky, 2012), we
1115 cannot rule out the possibility that more time is needed for post-encoding synaptic
1116 consolidation to become apparent at a functional level, or that that the cognitive demands
1117 induced by our distractor tasks were already sufficient to inhibit consolidation throughout
1118 the entire retention interval in both LS and SL conditions. That being said, recent EEG

1119 research suggests that the brain rapidly cycles between “online” and “offline” states during
1120 simple attention tasks, and that even ultra-short (i.e., seconds-long) bouts of offline time
1121 might already support the consolidation of previously encoded memories (Wamsley et al.,
1122 2023). More specifically, Wamsley et al. (2023) used a distractor task where participants
1123 were presented with a series of the digits 1 to 9, and were required to press a key as quickly
1124 as possible for each digit except one target digit. This task very much resembles our own
1125 distractor tasks from the current research; the former might even be considered more
1126 attentionally demanding because of the higher number of distractor items compared to our
1127 tasks (8 vs. 1). From this perspective, several minutes of our relatively easy distractor tasks
1128 should have been sufficient to allow for consolidation to occur. Future research should aim at
1129 further specifying the consolidation process by determining a critical post-encoding interval
1130 of reduced or minimized cognitive demands (see Mercer, 2015).

1131 Second, the effects of post-encoding consolidation might be more pronounced in
1132 delayed memory tests on time scales of hours or days. In our current research, memory was
1133 tested immediately after the 5-min retention interval. Some waking rest research suggests
1134 that consolidation effects might be especially pronounced after longer time intervals of one or
1135 more days (e.g., Martini et al., 2020), possibly because more consolidated memory traces are
1136 protected from non-specific retroactive interference that might build up during everyday
1137 activities outside the lab. To test such a mechanism with respect to the TGRI, future
1138 research might adapt the procedure from our current research to include additional memory
1139 tests delayed by several hours.

1140 Third, the facilitating effect of post-encoding consolidation might be more
1141 pronounced for learning materials such as paired associates that are more dependent on
1142 hippocampal resources than singletons as used in our current research. Indeed, the
1143 opportunistic consolidation theory explicitly refers to hippocampus-dependent memories
1144 (Mednick et al., 2011), and item memory (as opposed to associative memory) has been

1145 suggested to be more dependent on extrahippocampal medial temporal lobe structures (see
1146 Kuhlmann et al., 2019). Thus, it might be the case that in our current research, differential
1147 consolidation in LS and SL conditions was attenuated by our choice of learning materials.
1148 Future research might adapt our methodological approach for word pairs as learning
1149 material, for which suitable storage-retrieval MPT models are readily available (e.g., Riefer
1150 & Batchelder, 1995; Rouder & Batchelder, 1998).

1151 Further research is needed to more explicitly consider various methodological details
1152 that could be critical for consolidation effects to occur, such as longer retention intervals,
1153 delayed memory tests, and associative learning materials. Notwithstanding these open
1154 questions, future research might only add consolidation as an additional factor relevant to
1155 the TGRI under rather specific conditions, whereas a more comprehensive role for temporal
1156 distinctiveness seems very likely given our own findings and those by Ecker, Brown, et al.
1157 (2015).

1158 **Finding 3: The TGRI is not moderated by L1-L2 similarity**

1159 In Experiment 3, we extended our experimental design to include L1-L2 similarity as
1160 a second factor. We found that manipulating L2 items as words versus geometric figures did
1161 not affect any of our L1 memory measures. Although this result is in line with previous
1162 findings by Mercer (2015), it was very surprising because similarity is widely assumed to play
1163 a crucial role for retroactive interference. Indeed, early research found that higher L1-L2
1164 similarity leads to stronger retroactive interference effects in item memory (e.g., Johnson,
1165 1933; McGeoch & McDonald, 1931). For example, in their classic research, McGeoch and
1166 McDonald (1931) found that retroactive interference for previously learned adjectives was
1167 strongest with interpolated synonyms of the original items, and that it decreased steadily
1168 with antonyms, unrelated adjectives, nonsense syllables, and three-digit numbers. Later
1169 research, however, offered a more complex picture. More specifically, Dey (1969) found that
1170 on a continuum from low to high L1-L2 synonymy, a medium degree of synonymy resulted in

1171 the strongest retroactive interference, whereas effects were less pronounced for more or less
1172 synonymous items. Such a pattern is in line with the Skaggs-Robinson hypothesis of an
1173 inverted U-shaped relationship of L1-L2 similarity and retroactive interference (Robinson,
1174 1927; Skaggs, 1925). From such a perspective, it might have been the case that our
1175 operationalization of L1-L2 similarity resulted in two conditions that lie on opposite sides of
1176 a hypothetical Skaggs-Robinson function but have identical distances to its peak. Future
1177 research might scrutinize such a hypothesis by including more than just two L1-L2 similarity
1178 conditions.

1179 An alternative explanation for the absence of an L1-L2 similarity effect was offered by
1180 Mercer (2015). He observed that in his experiment, L2 recognition accuracies were
1181 significantly better for similar than for dissimilar L2 item pairs. Indeed, in our experiment,
1182 we observed the same pattern with respect to final L2 free recall accuracies. Mercer (2015)
1183 argued that higher cognitive demands might have been induced by the more difficult-to-learn
1184 item pairs in the low L1-L2 similarity condition, resulting in stronger L1 consolidation
1185 inhibition than in the high L1-L2 similarity condition. Thereby, the negative effect of higher
1186 L1-L2 similarity might have been equalized. While such an interpretation seems reasonable
1187 on the basis of surface memory measures, our MPT results provide evidence against it, since
1188 storage parameter s would have been expected to be higher in the high compared to the low
1189 L1-L2 similarity condition, which was not the case.

1190 Our null findings with respect to L1-L2 similarity are not in line with a generalized
1191 distinctiveness theory of retroactive interference (Brown et al., 2007; Ecker, Brown, et al.,
1192 2015). Thus, future research might instead consider alternative factors to test the
1193 assumption of a multidimensional memory space. One such factor might be the context
1194 within which L1 and L2 materials are presented. More specifically, previous research suggests
1195 that when L1 items are learned and retrieved in some context A, learning L2 items in a
1196 different context B (i.e., ABA) attenuates retroactive interference compared to a condition

1197 where all learning occurs within the same context A (i.e., AAA, see Shapiro & Levy-Gigi,
1198 2016). It would thus be interesting to see if L2 timing and L2 context show an interaction
1199 effect on L1 free recall accuracy and MPT recall retrieval as we predicted for L1-L2
1200 similarity in Experiment 3. For the time being, a purely *temporal* distinctiveness theory of
1201 the TGRI seems to capture the results from our current research best.

1202 **Limitations**

1203 Some limitations of the current research should be acknowledged. First, we used a
1204 modified version of the well-established storage-retrieval MPT model by Riefer and
1205 Batchelder (1995) to gain insights into the latent processes underlying the TGRI. While this
1206 model-based approach turned out to be very successful in our present application and can be
1207 viewed as a particular strength of the current research, future research should confirm the
1208 substantive interpretation of the model parameters by means of selective influence studies
1209 (see Schmidt et al., 2023). In our case, the model fit the data well across both Experiments 2
1210 and 3 regardless of specific equality constraints ($d = s * r_2$, $d = s$) and estimation approaches
1211 (maximum likelihood estimation for aggregated category frequencies, Bayesian hierarchical
1212 estimation for individual category frequencies, see also Erdfelder et al., in press; Singmann et
1213 al., 2024). Importantly, estimates for parameters s and r_1 were mirrored by our surface
1214 measures of free recall and recognition: Significant L2 timing effects on free recall along with
1215 unambiguous null effects on recognition already hint at a purely retrieval-driven effect (see
1216 Küpper-Tetzl & Erdfelder, 2012 for the same argument regarding free and cued recall). In
1217 this sense, our experiments can be seen as initial evidence for the validity of the model.

1218 Second, in line with Ecker, Brown, et al. (2015), we opted for a within-participants
1219 manipulation of L2 timing. This might have incentivized participants to develop certain
1220 strategies at encoding and to engage in active rehearsal during the retention interval. While
1221 our exploratory analysis of self-reported rehearsal in Experiment 3 confirmed that rehearsal
1222 was not a cause for the observed L2 timing effect, we only assessed rehearsal once in a

1223 post-experimental questionnaire. In principle, a trial-wise assessment of rehearsal would have
1224 allowed for a more fine-grained evaluation of rehearsal effects on our memory measures.
1225 However, repeatedly asking participants about their rehearsal engagement might have
1226 increased actual rehearsal above the already high levels we observed. Future research might
1227 either adopt a between-participants manipulation (including only one trial per participant)
1228 or opt for a subtle trial-wise rehearsal assessment.

1229 **Constraints on Generality**

1230 Our results provide strong evidence in favor of a retrieval-driven TGRI in adult
1231 participants. To approximate the sample characteristics reported by Ecker, Brown, et al.
1232 (2015), our two online samples from Experiments 1 and 2 were prescreened to only include
1233 native English speakers who were currently studying, whereas participants in Experiment 3
1234 were recruited within the premises of the University of Mannheim. As a result, our samples
1235 largely consisted of educated younger adults. Research on the effect of post-encoding waking
1236 rest suggests that, if anything, this effect might be even more pronounced in children and
1237 older adults (Martini et al., 2020), so we assume our findings to hold for healthy individuals
1238 of any age.

1239 The absence of any storage or consolidation contribution to the TGRI might in
1240 principle depend on the nature of the learning materials and the procedure. As detailed in
1241 the General Discussion section, longer L1-L2 and L2-T time intervals as well as associative
1242 learning material might provide more favorable conditions for a consolidation benefit to
1243 occur in LS compared to SL conditions.

1244 We have no reason to believe that the results depend on other characteristics of the
1245 participants, materials, or context.

1246 Conclusion

1247 Overall, our current research provides convincing evidence in line with a temporal
1248 distinctiveness account of the TGRI in free recall. Over a century after Müller and Pilzecker
1249 (1900) empirically demonstrated the TGRI for the first time, our results contribute to a
1250 better understanding of long-standing and so far unresolved issues in the literature on
1251 retroactive interference and consolidation in episodic memory. More research is needed to
1252 more explicitly identify the specific conditions under which consolidation processes might
1253 potentially add to an otherwise purely retrieval-driven memory phenomenon. We hope that
1254 our MPT analysis approach contributes to a more comprehensive inclusion of such cognitive
1255 modeling tools in the field.

References

1256

- 1257 Anderson, S. F., & Maxwell, S. E. (2016). There's more than one way to conduct a
1258 replication study: Beyond statistical significance. *Psychological Methods*, *21*(1), 1–12.
1259 <https://doi.org/10.1037/met0000051>
- 1260 Aust, F., & Barth, M. (2023). *Papaja: Prepare reproducible APA journal articles with R*
1261 *Markdown*.
- 1262 Berres, S., & Erdfelder, E. (2021). The sleep benefit in episodic memory: An integrative
1263 review and a meta-analysis. *Psychological Bulletin*, *147*(12), 1309–1353.
1264 <https://doi.org/10.1037/bul0000350>
- 1265 Brown, G. D. A., Neath, I., & Chater, N. (2007). A temporal ratio model of memory.
1266 *Psychological Review*, *114*(3), 539–576. <https://doi.org/10.1037/0033-295X.114.3.539>
- 1267 Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed). L. Erlbaum
1268 Associates.
- 1269 Craig, M., Ottaway, G., & Dewar, M. (2018). Rest on it: Awake quiescence facilitates insight.
1270 *Cortex*, *109*, 205–214. <https://doi.org/10.1016/j.cortex.2018.09.009>
- 1271 Delprato, D. J. (2005). Retroactive interference as a function of degree of interpolated study
1272 without overt retrieval practice. *Psychonomic Bulletin & Review*, *12*(2), 345–349.
1273 <https://doi.org/10.3758/BF03196383>
- 1274 Dewar, M., Cowan, N., & Della Sala, S. (2007). Forgetting due to retroactive interference: A
1275 fusion of Müller and Pilzecker's (1900) early insights into everyday forgetting and recent
1276 research on anterograde amnesia. *Cortex*, *43*(5), 616–634.
1277 [https://doi.org/10.1016/S0010-9452\(08\)70492-1](https://doi.org/10.1016/S0010-9452(08)70492-1)
- 1278 Dey, M. K. (1969). Retroactive inhibition as a function of similarity of meaning in free-recall
1279 learning. *Psychologische Forschung*, *33*(1), 79–84. <https://doi.org/10.1007/BF00424618>
- 1280 Dudai, Y. (2004). The neurobiology of consolidations, or, how stable is the engram? *Annual*
1281 *Review of Psychology*, *55*, 51–86.
1282 <https://doi.org/10.1146/annurev.psych.55.090902.142050>

- 1283 Ecker, U. K. H., Brown, G. D. A., & Lewandowsky, S. (2015). Memory without
1284 consolidation: Temporal distinctiveness explains retroactive interference. *Cognitive*
1285 *Science*, *39*(7), 1570–1593. <https://doi.org/10.1111/cogs.12214>
- 1286 Ecker, U. K. H., & Lewandowsky, S. (2012). Computational constraints in cognitive theories
1287 of forgetting. *Frontiers in Psychology*, *3*, 1–5. <https://doi.org/10.3389/fpsyg.2012.00400>
- 1288 Ecker, U. K. H., Tay, J.-X., & Brown, G. D. A. (2015). Effects of pre-study and post-study
1289 rest on memory: Support for temporal interference accounts of forgetting. *Psychonomic*
1290 *Bulletin & Review*, *22*(3). <https://doi.org/10.3758/s13423-014-0737-8>
- 1291 Erdfelder, E., Auer, T.-S., Hilbig, B. E., Aßfalg, A., Moshagen, M., & Nadarevic, L. (2009).
1292 Multinomial processing tree models: A review of the literature. *Zeitschrift für Psychologie*
1293 */ Journal of Psychology*, *217*(3), 108–124. <https://doi.org/10.1027/0044-3409.217.3.108>
- 1294 Erdfelder, E., Quevedo Pütter, J., & Schnuerch, M. (in press). On aggregation invariance of
1295 multinomial processing tree models. *Behavior Research Methods*.
- 1296 Faul, F., Erdfelder, E., Buchner, A., & Lang, A.-G. (2009). Statistical power analyses using
1297 G*Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods*,
1298 *41*(4), 1149–1160. <https://doi.org/10.3758/BRM.41.4.1149>
- 1299 Gelman, A., & Rubin, D. B. (1992). Inference from iterative simulation using multiple
1300 sequences. *Statistical Science*, *7*(4), 457–472. <https://doi.org/10.1214/ss/1177011136>
- 1301 Hager, W., & Hasselhorn, M. (Eds.). (1994). *Handbuch deutschsprachiger Wortnormen*
1302 *[Handbook of German-Language Word Norms]*. Hogrefe.
- 1303 Hartshorne, J. K., de Leeuw, J. R., Goodman, N. D., Jennings, M., & O'Donnell, T. J.
1304 (2019). A thousand studies for the price of one: Accelerating psychological science with
1305 Pushkin. *Behavior Research Methods*, *51*(4), 1782–1803.
1306 <https://doi.org/10.3758/s13428-018-1155-z>
- 1307 Heck, D. W., Arnold, N. R., & Arnold, D. (2018). TreeBUGS: An R package for hierarchical
1308 multinomial-processing-tree modeling. *Behavior Research Methods*, *50*(1), 264–284.
1309 <https://doi.org/10.3758/s13428-017-0869-7>

- 1310 Henninger, F., Shevchenko, Y., Mertens, U. K., Kieslich, P. J., & Hilbig, B. E. (2022).
1311 Lab.js: A free, open, online study builder. *Behavior Research Methods*, *54*(2), 556–573.
1312 <https://doi.org/10.3758/s13428-019-01283-5>
- 1313 Humiston, G. B., & Wamsley, E. J. (2018). A brief period of eyes-closed rest enhances motor
1314 skill consolidation. *Neurobiology of Learning and Memory*, *155*, 1–6.
1315 <https://doi.org/10.1016/j.nlm.2018.06.002>
- 1316 Johnson, L. M. (1933). Similarity of meaning as a factor in retroactive inhibition. *The*
1317 *Journal of General Psychology*, *9*(2), 377–389.
1318 <https://doi.org/10.1080/00221309.1933.9920942>
- 1319 King, O., & Nicosia, J. (2022). The effects of wakeful rest on memory consolidation in an
1320 online memory study. *Frontiers in Psychology*, *13*, 1–13.
1321 <https://doi.org/10.3389/fpsyg.2022.932592>
- 1322 Klauer, K. C. (2010). Hierarchical multinomial processing tree models: A latent-trait
1323 approach. *Psychometrika*, *75*(1), 70–98. <https://doi.org/10.1007/s11336-009-9141-0>
- 1324 Kuhlmann, B. G., Erdfelder, E., & Moshagen, M. (2019). Testing Interactions in
1325 Multinomial Processing Tree Models. *Frontiers in Psychology*, *10*, 1–11.
1326 <https://doi.org/10.3389/fpsyg.2019.02364>
- 1327 Küpper-Tetzl, C. E., & Erdfelder, E. (2012). Encoding, maintenance, and retrieval
1328 processes in the lag effect: A multinomial processing tree analysis. *Memory*, *20*(1), 37–47.
1329 <https://doi.org/10.1080/09658211.2011.631550>
- 1330 Lakens, D. (2013). Calculating and reporting effect sizes to facilitate cumulative science: A
1331 practical primer for t-tests and ANOVAs. *Frontiers in Psychology*, *4*, 1–12.
1332 <https://doi.org/10.3389/fpsyg.2013.00863>
- 1333 Landauer, T. K. (1974). Consolidation in human memory: Retrograde amnesic effects of
1334 confusable items in paired-associate learning. *Journal of Verbal Learning and Verbal*
1335 *Behavior*, *13*(1), 45–53. [https://doi.org/10.1016/S0022-5371\(74\)80029-0](https://doi.org/10.1016/S0022-5371(74)80029-0)

- 1336 Lange, K., Kühn, S., & Filevich, E. (2015). "Just Another Tool for Online Studies"
1337 (JATOS): An easy solution for setup and management of web servers supporting online
1338 studies. *PLOS ONE*, *10*(6), 1–14. <https://doi.org/10.1371/journal.pone.0130834>
- 1339 Leetham, E., Watermeyer, T., & Craig, M. (2024). An online experiment that presents
1340 challenges for translating rest-related gains in visual detail memory from the laboratory
1341 to naturalistic settings. *PLOS ONE*, *19*(1), 1–24.
1342 <https://doi.org/10.1371/journal.pone.0290811>
- 1343 Leys, C., Ley, C., Klein, O., Bernard, P., & Licata, L. (2013). Detecting outliers: Do not use
1344 standard deviation around the mean, use absolute deviation around the median. *Journal*
1345 *of Experimental Social Psychology*, *49*(4), 764–766.
1346 <https://doi.org/10.1016/j.jesp.2013.03.013>
- 1347 Martini, M., Heinz, A., Hinterholzer, J., Martini, C., & Sachse, P. (2020). Effects of wakeful
1348 resting versus social media usage after learning on the retention of new memories.
1349 *Applied Cognitive Psychology*, *34*(2), 551–558. <https://doi.org/10.1002/acp.3641>
- 1350 Martini, M., & Sachse, P. (2020). Factors modulating the effects of waking rest on memory.
1351 *Cognitive Processing*, *21*(1), 149–153. <https://doi.org/10.1007/s10339-019-00942-x>
- 1352 Mathôt, S., Schreij, D., & Theeuwes, J. (2012). OpenSesame: An open-source, graphical
1353 experiment builder for the social sciences. *Behavior Research Methods*, *44*(2), 314–324.
1354 <https://doi.org/10.3758/s13428-011-0168-7>
- 1355 McClelland, J. L., McNaughton, B. L., & O'Reilly, R. C. (1995). Why there are
1356 complementary learning systems in the hippocampus and neocortex: Insights from the
1357 successes and failures of connectionist models of learning and memory. *Psychological*
1358 *Review*, *102*(3), 419–457. <https://doi.org/10.1037/0033-295X.102.3.419>
- 1359 McGaugh, J. L. (2000). Memory—a century of consolidation. *Science*, *287*(5451), 248–251.
1360 <https://doi.org/10.1126/science.287.5451.248>

- 1361 McGeoch, J. A. (1933). Studies in retroactive inhibition: II. Relationships between temporal
1362 point of interpolation, length of interval, and amount of retroactive inhibition. *The*
1363 *Journal of General Psychology*, *9*(1), 44–57.
1364 <https://doi.org/10.1080/00221309.1933.9920912>
- 1365 McGeoch, J. A., & McDonald, W. T. (1931). Meaningful relation and retroactive inhibition.
1366 *The American Journal of Psychology*, *43*(4), 579–588. <https://doi.org/10.2307/1415159>
- 1367 Mednick, S. C., Cai, D. J., Shuman, T., Anagnostaras, S., & Wixted, J. T. (2011). An
1368 opportunistic theory of cellular and systems consolidation. *Trends in Neurosciences*,
1369 *34*(10), 504–514. <https://doi.org/10.1016/j.tins.2011.06.003>
- 1370 Mercer, T. (2015). Wakeful rest alleviates interference-based forgetting. *Memory*, *23*(2),
1371 127–137. <https://doi.org/10.1080/09658211.2013.872279>
- 1372 Moshagen, M. (2010). multiTree: A computer program for the analysis of multinomial
1373 processing tree models. *Behavior Research Methods*, *42*(1), 42–54.
1374 <https://doi.org/10.3758/BRM.42.1.42>
- 1375 Müller, G. E., & Pilzecker, A. (1900). Experimentelle Beiträge zur Lehre vom Gedächtniss
1376 [Experimental contributions to the science of memory]. *Zeitschrift für Psychologie*,
1377 *Ergänzungsband [Supplement] 1*, 1–300.
- 1378 Nadarevic, L. (2017). Emotionally enhanced memory for negatively arousing words: Storage
1379 or retrieval advantage? *Cognition & Emotion*, *31*(8), 1557–1570.
1380 <https://doi.org/10.1080/02699931.2016.1242477>
- 1381 Riefer, D. M., & Batchelder, W. H. (1995). A multinomial modeling analysis of the
1382 recognition-failure paradigm. *Memory & Cognition*, *23*(5), 611–630.
1383 <https://doi.org/10.3758/BF03197263>
- 1384 Robinson, E. S. (1920). Some factors determining the degree of retroactive inhibition.
1385 *Psychological Monographs*, *28*(6), 1–57. <https://doi.org/10.1037/h0093155>
- 1386 Robinson, E. S. (1927). The 'similarity' factor in retroaction. *The American Journal of*
1387 *Psychology*, *39*(1), 297–312. <https://doi.org/10.2307/1415419>

- 1388 Rouder, J. N., & Batchelder, W. H. (1998). Multinomial models for measuring storage and
1389 retrieval processes in paired associate learning. In C. E. Dowling, F. S. Roberts, & P.
1390 Theuns (Eds.), *Recent Progress in Mathematical Psychology* (pp. 195–226). Psychology
1391 Press.
- 1392 Schmidt, O., Erdfelder, E., & Heck, D. W. (2023). How to develop, test, and extend
1393 multinomial processing tree models: A tutorial. *Psychological Methods*. Advance online
1394 publication. <https://doi.org/10.1037/met0000561>
- 1395 Schonfield, D., & Robertson, B. A. (1966). Memory storage and aging. *Canadian Journal of*
1396 *Psychology / Revue Canadienne de Psychologie*, *20*(2), 228–236.
1397 <https://doi.org/10.1037/h0082941>
- 1398 Shapiro, A. R., & Levy-Gigi, E. (2016). Susceptibility to retroactive interference: The effect
1399 of context as a function of age and cognition. *Memory*, *24*(3), 399–408.
1400 <https://doi.org/10.1080/09658211.2015.1011168>
- 1401 Singmann, H., Heck, D. W., Barth, M., Erdfelder, E., Arnold, N. R., Aust, F., Calanchini, J.,
1402 Gümüşdaglı, F. E., Horn, S. S., Kellen, D., Klauer, K. C., Matzke, D., Meissner, F.,
1403 Michalkiewicz, M., Schaper, M. L., Stahl, C., Kuhlmann, B. G., & Groß, J. (2024).
1404 Evaluating the robustness of parameter estimates in cognitive models: A meta-analytic
1405 review of multinomial processing tree models across the multiverse of estimation methods.
1406 *Psychological Bulletin*. Advance online publication. <https://doi.org/10.1037/bul0000434>
- 1407 Singmann, H., & Kellen, D. (2013). MPTinR: Analysis of multinomial processing tree
1408 models in R. *Behavior Research Methods*, *45*(2), 560–575.
1409 <https://doi.org/10.3758/s13428-012-0259-0>
- 1410 Skaggs, E. B. (1925). Further studies in retroactive inhibition. *Psychology Monograph*, *34*(8),
1411 1–60.

- 1412 Varma, S., Takashima, A., Krewinkel, S., van Kooten, M., Fu, L., Medendorp, W. P.,
1413 Kessels, R. P. C., & Daselaar, S. M. (2017). Non-interfering effects of active
1414 post-encoding tasks on episodic memory consolidation in humans. *Frontiers in*
1415 *Behavioral Neuroscience*, *11*, Article 54. <https://doi.org/10.3389/fnbeh.2017.00054>
- 1416 Wamsley, E. J. (2019). Memory consolidation during waking rest. *Trends in Cognitive*
1417 *Sciences*, *23*(3), 171–173. <https://doi.org/10.1016/j.tics.2018.12.007>
- 1418 Wamsley, E. J., Arora, M., Gibson, H., Powell, P., & Collins, M. (2023). Memory
1419 consolidation during ultra-short offline states. *Journal of Cognitive Neuroscience*, *35*(10),
1420 1617–1634. https://doi.org/10.1162/jocn_a_02035
- 1421 Wickelgren, W. A. (1974). Single-trace fragility theory of memory dynamics. *Memory &*
1422 *Cognition*, *2*(4), 775–780. <https://doi.org/10.3758/BF03198154>
- 1423 Wickelgren, W. A. (1977). *Learning and memory*. Prentice-Hall.
- 1424 Wilson, M. (1988). MRC psycholinguistic database: Machine-usable dictionary, version 2.00.
1425 *Behavior Research Methods*, *20*(1), 6–10. <https://doi.org/10.3758/BF03202594>
- 1426 Wixted, J. T. (2004). The psychology and neuroscience of forgetting. *Annual Review of*
1427 *Psychology*, *55*, 235–269. <https://doi.org/10.1146/annurev.psych.55.090902.141555>
- 1428 Wixted, J. T. (2010). The role of retroactive interference and consolidation in everyday
1429 forgetting. In S. Della Sala (Ed.), *Forgetting* (pp. 285–312). Psychology Press.

Appendix A

Results of the Preregistered Storage-Retrieval MPT Analysis of Experiment 2

Table A1

Results of the Preregistered Storage-Retrieval MPT Analysis of Experiment 2 with Restriction

$$d = s * r_2$$

Parameter	LS	SL
Aggregated data ^a		
s	.63 [.61, .65]	.64 [.62, .66]
r_1	.62 [.59, .65]	.58 [.55, .61]
r_2	.93 [.92, .95]	.93 [.91, .94]
g	.38 [.37, .40]	.38 [.37, .40]
Individual data ^b		
s	.66 [.61, .70]	.68 [.63, .72]
r_1	.63 [.59, .68]	.57 [.53, .62]
r_2	.97 [.95, .99]	.96 [.94, .98]
g	.32 [.29, .36]	.32 [.29, .36]

Note. LS = Long-Short (i.e., long L1-L2 interval, short L2-T interval), SL = Short-Long (i.e., short L1-L2 interval, long L2-T interval). Parameter s = probability of storing an L1 word, r_1 = probability of retrieving an L1 word during recall, r_2 = probability of retrieving an L1 word during recognition, g = probability of guessing 'old' during recognition. Parameter g was restricted to be equal between conditions. The model fit the data well, both for the aggregated data, $G^2(1) = 1.47$, $p = .225$, and for the individual data, $p_1 = .461$, $p_2 = .306$. 95% confidence intervals (for the aggregated data) or Bayesian credibility intervals (for the individual data) are indicated in brackets.

Appendix B

A Priori MPT Power Analysis for Experiment 3

1430 To compute an a priori multinomial processing tree (MPT) power analysis in multiTree,
1431 “true” population values of the parameters under the H_1 model, and a to-be-tested H_0 model
1432 need to be specified (see Moshagen, 2010). We specified expected population values based on
1433 the corresponding parameter estimates from Experiment 2, considering that our switch from
1434 an online to a lab setting might generally increase memory performance. The exact values
1435 are provided in Table B1. We chose an H_1 model with perfect fit to the expected data (i.e.,
1436 $g_2 = g_1, g_4 = g_3^3$), and an H_0 model with additional constraints defining the hypothesis of a
1437 null interaction effect of L2 timing and L1-L2 similarity on recall retrieval parameter r_1 (i.e.,
1438 $\alpha_{LS} = \alpha_{SL}$). This specification lead to a required number of observations of 20,777. As each
1439 participant in Experiment 3 would be presented with a total of 160 relevant items (10 target
1440 and 10 distractor words per trial, 8 trials per participant), this number translated to a
1441 required sample size of $N = 20,777/160 \approx 130$.

³ Subscript 1 corresponds to "LS-high L1-L2 similarity" condition, subscript 2 to "SL-high L1-L2 similarity" condition, subscript 3 to "LS-low L1-L2 similarity" condition, and subscript 4 to "SL-low L1-L2 similarity" condition.

Table B1*Power Analysis Population Specifications for Experiment 3*

Parameter	High similarity		Low similarity	
	LS	SL	LS	SL
s	.70	.70	.70	.70
r_1	.68 ^a	.60 ^a	.80	.80
r_2	.98	.98	.98	.98
g	.40	.40	.40	.40

Note. LS = Long-Short (i.e., long L1-L2 interval, short L2-T interval), SL = Short-Long (i.e., short L1-L2 interval, long L2-T interval). Parameter s = probability of storing an L1 word, r_1 = probability of retrieving an L1 word during recall, r_2 = probability of retrieving an L1 word during recognition, g = probability of guessing 'old' during recognition.

^a Parameter r_1 values in both high L1-L2 similarity conditions are implied by shrinkage parameters $\alpha_{LS} = .85$ and $\alpha_{SL} = .75$, that is, $r_{11} = r_{13} * \alpha_{LS}$, $r_{12} = r_{14} * \alpha_{SL}$

Appendix C

Results of the Storage-Retrieval MPT Sensitivity Analysis of Experiment 3

Table C1

Results of the Storage-Retrieval MPT Sensitivity Analysis of Experiment 3

Parameter	High similarity		Low similarity	
	LS	SL	LS	SL
Aggregated data				
s	.84 [.82, .86]	.84 [.82, .86]	.85 [.84, .87]	.84 [.82, .85]
r_1	.68 [.66, .71]	.62 [.59, .65]	.65 [.62, .67]	.65 [.62, .67]
r_2	.98 [.98, .99]	.99 [.98, .99]	.99 [.99, .99]	.99 [.99, .99]
g	.36 [.33, .39]	.36 [.33, .39]	.36 [.33, .39]	.36 [.33, .39]
Individual data				
s	.88 [.85, .91]	.87 [.84, .89]	.88 [.85, .90]	.86 [.83, .89]
r_1	.68 [.63, .74]	.63 [.58, .67]	.66 [.60, .72]	.65 [.60, .70]
r_2	.99 [.98, .99]	.99 [.98, .99]	.99 [.99, .99]	.99 [.99, .99]
g	.32 [.26, .39]	.32 [.26, .39]	.32 [.26, .39]	.32 [.26, .39]

Note. LS = Long-Short (i.e., long L1-L2 interval, short L2-T interval), SL = Short-Long (i.e., short L1-L2 interval, long L2-T interval). Parameter s = probability of storing an L1 word, r_1 = probability of retrieving an L1 word during recall, r_2 = probability of retrieving an L1 word during recognition, g = probability of guessing 'old' during recognition. Parameter g was restricted to be equal between conditions. 95% confidence intervals (for the aggregated data) or Bayesian credibility intervals (for the individual data) are indicated in brackets.

Waking rest during retention facilitates memory consolidation, but so does social media use: A model-based storage-retrieval analysis

Julian Quevedo Pütter^{1,*} and Edgar Erdfelder¹

¹University of Mannheim, School of Social Sciences, Mannheim, Germany

*julian@quevedo.de

ABSTRACT

A short period of post-encoding waking rest has been shown to benefit subsequent memory performance. For example, past research suggests that waking rest after learning Icelandic-German word pairs boosts subsequent recall relative to an equally long period of social media use. Such findings are typically interpreted as evidence in favor of *diversion retroactive interference*. According to this account, non-specific cognitive processing inhibits consolidation and thus impairs *storage* of information encoded previously. However, the effect might alternatively be explained by *similarity retroactive interference* according to which *retrieval* is hampered by information processed during retention. Here, we report two experiments that shed light on the mechanisms underlying the waking rest effect. In both experiments, participants either wakefully rested, used social media, or engaged in additional Norwegian-German vocabulary learning after the original learning phase. We performed multinomial processing tree (MPT) analyses to disentangle latent storage and retrieval contributions to cued recall and recognition performance. We did not find any memory differences between the waking rest and social media conditions in either experiment. Moreover, storage, but not retrieval, was reliably impaired in the vocabulary condition. Thereby, the present research provides direct behavioral evidence for a dominant role of consolidation in the waking rest effect.

Introduction

Waking rest has been defined as a period of quiet, reflective thought void of distracting stimuli¹. Superficially, such resting periods might appear rather unproductive or even a waste of time. However, quite to the contrary, waking rest has not only been linked to mental health and sleep benefits¹, but is also argued to facilitate memory consolidation². Indeed, a growing body of evidence suggests that wakefully resting after new learning can enhance subsequent memory performance compared to engaging in a cognitively demanding distractor task²⁻⁴.

In a recent study by Martini et al.⁵, social media use after learning new vocabulary has been shown to be detrimental to subsequent memory performance relative to a waking rest condition. More specifically, participants in this study learned and immediately recalled Icelandic-German word pairs, before being randomly assigned to either a waking rest or a social media condition. In the waking rest condition, participants rested for 8 minutes, whereas in the social media condition, participants used Facebook or Instagram for the same amount of time. In two delayed recall tests immediately after the 8-min retention interval and again after 24 hours, participants in the waking rest condition showed significantly less forgetting relative to the immediate recall than participants in the social media condition.

Given its simplicity, waking rest might be a promising behavioral intervention for improving memory in many applied settings⁶⁻⁸. However, such an optimistic perspective is challenged by a considerable number of studies that have failed to find significant effects of waking rest⁹. Thus, our first aim in the present research was to conduct a close replication of the social media study reported by Martini et al.⁵. In our opinion, this study represents a particularly important replication target¹⁰ because of the combination of high practical relevance and rather low experimental control (i.e., largely unrestricted social media use) compared to other tasks that have been used before (e.g., spot-the-difference task¹¹, further word pair learning¹²).

On a theoretical level, waking rest has been ascribed a central role in memory consolidation during wakefulness. According to the opportunistic theory of memory consolidation¹³, waking rest facilitates the consolidation of recently acquired memories by protecting limited hippocampal resources from retroactive interference. Conversely, any kind of distractor task that induces some minimal degree of cognitive processing or encoding demands should reduce the resources available for memory consolidation. Critically, this reasoning ignores possible contributions from similarity-based retroactive interference that might arise from similarities between the original learning material and the distractor task¹⁴. Indeed, Dewar et al.³ proposed an elegant theoretical model that differentiates two different types of retroactive interference: diversion retroactive interference

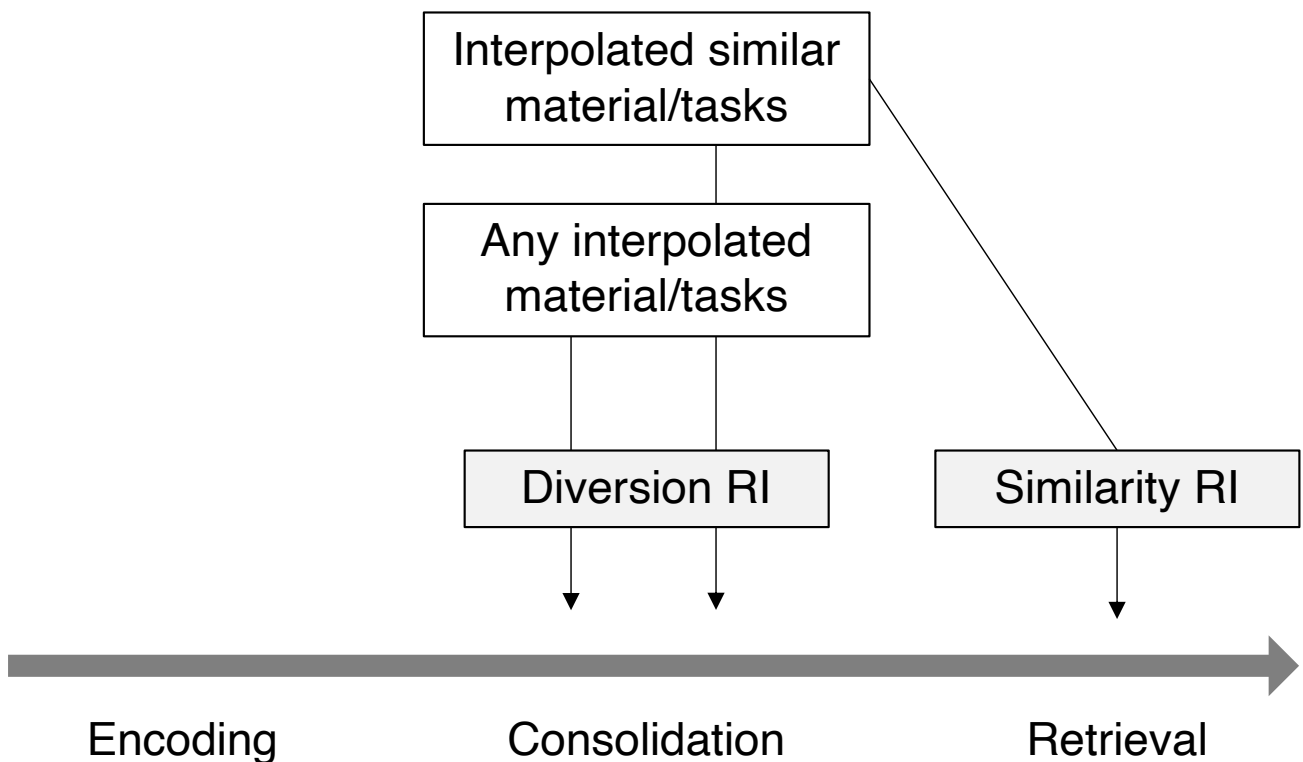


Figure 1. Illustration of the Dewar et al. theoretical model of retroactive interference³. Two types of retroactive interference are differentiated: diversion and similarity retroactive interference. First, any interpolated material or task will induce diversion retroactive interference and thereby interfere with memory consolidation. Second, only interpolated similar material and tasks will additionally induce similarity retroactive interference and thereby impair memory retrieval.

and similarity retroactive interference (see Figure 1). Whereas diversion retroactive interference is assumed to be induced by any interpolated task or material and to inhibit consolidation, similarity retroactive interference is only induced by similar tasks or material and impairs retrieval of the target information due to reduced discriminability from the interfering information.

It follows that positive effects of post-encoding waking rest might represent an unknown combination of both diversion and similarity retroactive interference. In other words, differences between waking rest and distractor conditions that have been interpreted as evidence in favor of opportunistic consolidation might in many cases just as well be explained solely by retrieval differences or some unknown combination of both consolidation and retrieval processes. For example, in the study by Martini et al.⁵, participants in the social media condition might have engaged in posts that were semantically related to the previously studied word pairs. Thus, strictly speaking, there is currently no direct behavioral evidence for a role of consolidation in the waking rest effect.

Our second aim in the present research was to close this critical gap in the literature by using multinomial processing tree (MPT) modeling to precisely disentangle consolidation and retrieval contributions to memory performance^{15,16}. Over the past decades, MPT models have been successfully applied in many different areas of psychology¹⁶. Storage-retrieval models represent a subset of MPT models that allow for disentangling storage and retrieval contributions to performances in some memory testing procedures.

A storage-retrieval MPT model that is ideally suited for the paradigm used by Martini et al.⁵ was developed by Riefer and Batchelder¹⁷. It is tailored to a recognition-then-cued-recall paradigm for word pairs, that is, it allows for Icelandic-German vocabulary as learning material and only requires the inclusion of an additional old-new recognition test for the German target words (presented without their Icelandic cue words). Thereby, a given target word might fall into one of four possible response categories: successful recognition and cued recall (Rn+ Rc+), successful recognition and unsuccessful cued recall (Rn+ Rc-), unsuccessful recognition and successful cued recall (Rn- Rc+), or unsuccessful recognition and cued recall (Rn- Rc-). The probabilities of these response categories (which can be estimated from their observed frequencies) are reparameterized by means of a set of latent model parameters. Each parameter represents the probability of some cognitive processing step: First,

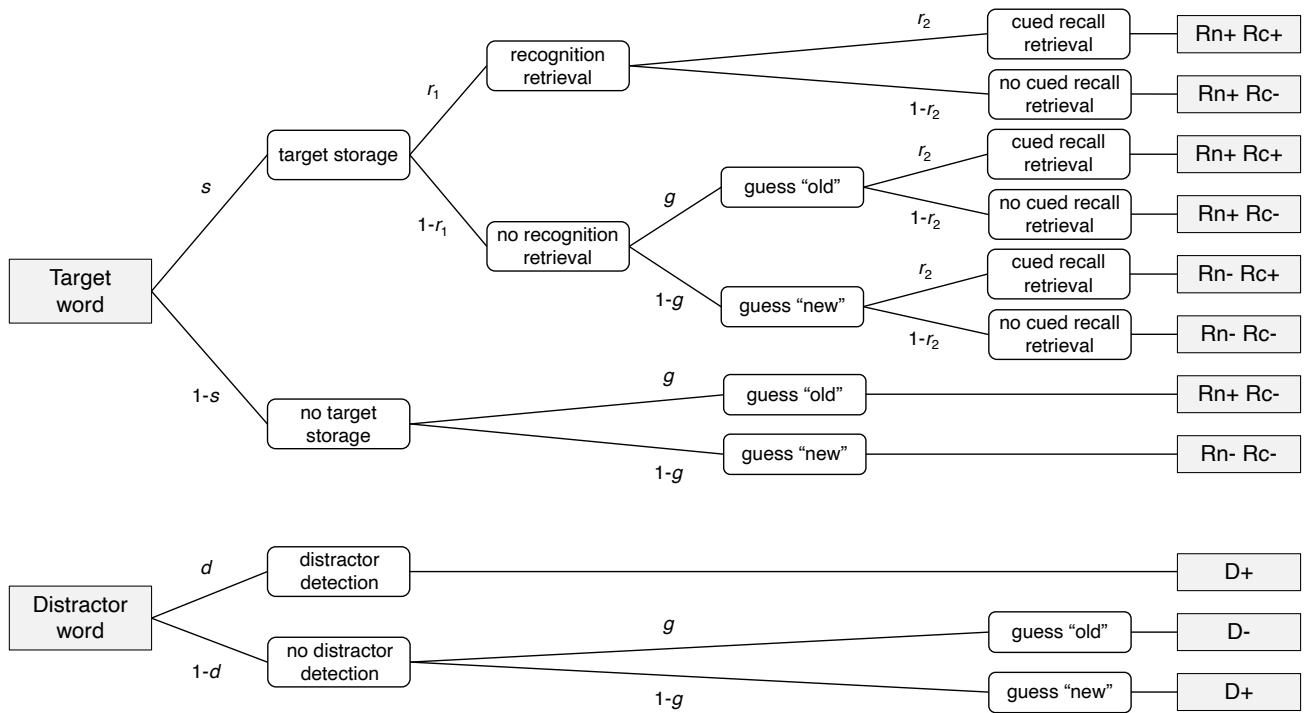


Figure 2. Illustration of the recognition-then-cued-recall storage-retrieval multinomial processing tree (MPT) model by Riefer and Batchelder^{17,18}. Each branch of the processing tree represents one possible sequence of cognitive processes, reflected in the following parameters: s = probability of successful target storage, r_1 = probability of successful target retrieval during recognition, g = probability of guessing ‘old’ during recognition, r_2 = probability of successful target retrieval during recall, d = probability of successful distractor detection during recognition. For a given target word, responses in the recognition and cued recall tests can be categorized as follows: Rn+ Rc+ = successful recognition and recall, Rn+ Rc- = successful recognition and unsuccessful recall, Rn- Rc+ = unsuccessful recognition and successful recall, Rn- Rc- = unsuccessful recognition and recall. For a given distractor word during recognition, D+ = correct rejection, D- = false alarm. Parameter d can be set equal to $s * r_1$ ¹⁸.

a given target word will be successfully stored with probability s , whereas storage will fail with probability $1 - s$. Next, a successfully stored target word will be retrieved versus not be retrieved during recognition with probabilities r_1 versus $1 - r_1$, respectively. When retrieval during recognition fails, or a target word was not successfully stored in the first place, a participant will correctly guess ‘old’ versus incorrectly guess ‘new’ with probabilities g versus $1 - g$, respectively. Finally, a successfully stored target word will be retrieved versus not be retrieved during cued recall with probabilities r_2 versus $1 - r_2$, respectively. The full tree structure of this model is illustrated in Figure 2.

Parameters s and r_2 of this MPT model can be related to diversion and similarity retroactive interference in a straightforward manner. First, parameter s represents the probability of a target word being stored in memory across the retention interval. Thereby, it represents a combination of successful encoding and consolidation contributions. Because the experimental manipulation in waking rest studies takes place only after the learning phase, it cannot affect encoding. Thus, differences in parameter s between conditions mirror differences in storage due to diversion retroactive interference. Second, parameter r_2 represents the probability of a stored target word being retrievable during cued recall. Accordingly, it should reflect effects of similarity retroactive interference. Note that the recognition retrieval parameter r_1 is not of substantive interest here. In fact, the probability of retrieving a stored target in an old-new recognition test should always be close to 1.

The model version originally proposed by Riefer and Batchelder¹⁷ can be adapted for our current purpose. First, memory performance for distractor items in the recognition test can be included into the model¹⁸. A distractor word always falls in one of two possible response categories: correct rejection (D+) or false alarm (D-). Parameter d represents the probability of successfully detecting a distractor word during recognition. By assuming this probability to be equal to the probability of successfully retrieving a target word during recognition (i.e., $d = s * r_1$, as proposed by Nadarevic¹⁸), inclusion of distractor words into the model yields a saturated model version (i.e., $df = 0$) with equal numbers of four non-redundant category

frequencies ($4 - 1 = 3$ for the target words plus $2 - 1 = 1$ for the distractor words) and four to-be-estimated parameters (s , r_1 , g , r_2). By introducing further equality constraints between experimental conditions, testable model versions with $df > 0$ can be defined that allow for model fit evaluations within a maximum likelihood (ML) framework¹⁹.

Second, the original recognition-then-cued-recall procedure can be replaced by a reversed cued-recall-then-recognition test sequence. The corresponding cued-recall-then-recognition storage-retrieval MPT model is mathematically equivalent to the original model but remedies the potential problem of the original procedure that cued recall performance might be positively biased by target presentation in the preceding old-new recognition test.

In the present research, we conducted two lab experiments to evaluate the replicability of the waking rest versus social media effect found by Martini and collaborators⁵, and to disentangle the contributions of diversion and similarity retroactive interference mediated by storage and retrieval effects, respectively. Both studies mainly differed in the order of memory tests, that is, Experiment 1 involved the original recognition-then-cued-recall testing procedure, whereas Experiment 2 involved a reversed cued-recall-then-recognition procedure.

In both studies, we extended the original study design by including a vocabulary condition in which participants engaged in an intentional learning task¹². Thereby, our design included not only a low similarity (i.e., social media), but also a high similarity (i.e., vocabulary) distractor condition. Thus, based on the Dewar et al. theoretical model³, we expected similarity retroactive interference to be lowest in the waking rest and highest in the vocabulary condition. In addition, we expected diversion retroactive interference to be lowest in the waking rest condition, but not to differ necessarily between the social media and the vocabulary condition. Our hypotheses, study protocols, and analysis plans for both studies were preregistered on the Open Science Framework (OSF)^{20,21}.

Experiment 1

Methods

Participants

We conducted an a priori power analysis in G*Power²² with the aim to obtain a statistical power of $1 - \beta = 80\%$ to detect a medium effect size of Cohen's $f = 0.25$ with $\alpha = 5\%$ in a one-factorial between-participants ANOVA model with three experimental conditions. This setup yielded a required sample size of $N = 159$ participants (i.e., $n = 53$ per condition).

Participants were recruited at the University of Mannheim. Interested individuals were eligible for participation only if (a) their first language was German or they were fluent in German, (b) they were between 18 and 32 years old (this corresponds to the age range in the sample from the original Martini et al.⁵ study), (c) they did not know any Icelandic or Norwegian, and (d) they had a smartphone with the Instagram app installed. Participants were randomly assigned to one of the experimental conditions using block randomization to ensure equal group sizes. All participants received course credit for an estimated net study duration of 45 minutes. In line with the original report by Martini et al., no further exclusion criteria were preregistered.

We collected data from 159 participants. Mean age in the full sample was 22.74 years ($SD = 2.50$, $range = 18-31$). 109 participants (68.55%) indicated to be female, 49 participants (30.82%) indicated to be male, and one participant identified their gender as non-binary. 154 participants (96.86%) indicated German to be their first language, and all participants confirmed to be fluent in German. 143 participants (89.94%) were studying at the time of participation, with 99 participants (62.26%) being enrolled in a psychology program.

The study was conducted in accordance with the Declaration of Helsinki (2013). As the study did not involve deception or other ethically relevant elements, formal approval from the ethics committee was not necessary according to the regulations of the ethics committee of the University of Mannheim. Informed consent was obtained from all participants.

Design

We used a between-participants manipulation of post-encoding activity with three levels: waking rest (minimal diversion and similarity retroactive interference), social media (high diversion and low similarity retroactive interference), and vocabulary (high diversion and similarity retroactive interference).

Material

To generate Icelandic-German word pairs for the original learning phase, Norwegian-German word pairs for the interpolated learning phase in the vocabulary condition, and German distractor words for the first and second delayed recognition tests, 84 German 5-letter nouns were selected from a word pool provided by Dimigen et al.²³. For the 24 Icelandic-German word pairs, German words were chosen such that their Icelandic translations did not resemble their German or English translations. For the 20 Norwegian-German word pairs, the same rule was applied. We opted against using the original Martini et al.⁵ Icelandic-German word pairs due to the necessity to generate an additional pool of 20 distractor words per recognition test that shared the characteristics of the target words.

Procedure

The procedure of our study included two experimental sessions separated by about 24 hours. Both sessions took place in the same laboratory room equipped with movable walls between work stations. Participants were told that our research aim was to investigate leisure activities of university students.

Session 1 First, participants provided informed consent and were checked for inclusion criteria. To comply as closely as possible with the original procedure, participants were then asked to report their current arousal and valence levels on a 7-point Likert scale. Next, during the learning phase, participants were presented with 24 Icelandic-German word pairs. They were told to memorize the material for a vocabulary test that would follow immediately. Each word pair was presented for 12 seconds and with a 3-sec inter-stimulus interval (ISI) in white font on a black background. Whereas the first 20 word pairs were presented in randomized order, the last 4 word pairs served as buffer items, that is, they were always presented in the same order and were not included in any memory tests. For the immediate cued recall, all 20 Icelandic cue words were presented simultaneously on the computer screen. Participants were given 3 minutes to type in as many of the previously learned German target words as possible. In contrast to the original procedure, we applied a learning criterion of 35% (i.e., at least 7 correct responses) to ensure sufficiently high frequencies in all response categories of the storage-retrieval MPT model. If a participant failed to reach this learning criterion during the first study-test cycle, they repeated this part of the procedure up to two additional times. Participants who did not reach the learning criterion in any repetition were excluded from further participation.

Next, an 8-min retention interval followed, during which participants engaged in their respective post-encoding activity. Participants in the waking rest and social media conditions were provided with headphones to minimize acoustic distractions. In the waking rest condition, they were instructed to relax as much as possible, but not to fall asleep. They were asked to lay their heads on their arms and to close their eyes. In the social media condition, participants were asked to engage in as many Instagram posts as possible on their own smartphones, but not to submit any own posts and not to follow any external links. They were also asked to use Instagram without tone to not distract other participants. During the entire experimental procedure, the shutters were closed and the lights turned off for all conditions. After 8 minutes had passed, participants in the waking rest and social media conditions received an acoustic signal over the headphones. In the vocabulary condition, participants learned and immediately recalled 20 Norwegian-German word pairs for 8 minutes. None of the German target words were previously included in the original learning phase. The procedure was the same as in the original learning phase, but there were no buffer items and no learning criterion was used.

After their respective post-encoding activity, all participants were again asked for their current arousal and valence levels. Afterwards, participants engaged in a first surprise delayed recognition-then-cued-recall test procedure. In the recognition test, participants were presented with a randomized sequence of 40 German words (i.e., 20 'old' target and 20 'new' distractor words). They were asked to indicate for each word whether it was 'old' (i.e., previously presented as part of an Icelandic-German word pair) by pressing the 'S' key on their keyboard or 'new' (i.e., not previously presented) by pressing the 'L' key. No time limit was imposed during recognition. The first delayed cued recall test was identical to the immediate cued recall test.

To conclude Session 1, participants were instructed to answer questions concerning thoughts about and conscious rehearsal of the Icelandic-German word pairs during the 8-min retention interval. In the social media condition, participants were additionally asked to estimate the total number of Instagram posts and the number of Instagram posts in Icelandic language they had engaged in.

Session 2 The procedure of Session 2 encompassed a second surprise delayed recognition-then-cued-recall test procedure that was identical to the first delayed test procedure from Session 1, except for a new set of 20 distractor words in the recognition test. Finally, participants were asked for information concerning the 24-hr interval between sessions, including sleep times, alcohol consumption, and the same questions concerning thoughts about and active rehearsal of the Icelandic vocabulary as in Session 1. After providing demographic information, participants were thanked and debriefed.

Data analysis

All analyses were conducted in R²⁴. A significance level of $\alpha = 5\%$ was used for all analyses.

We computed cued recall retention and recognition performance scores as dependent variables. For the cued recall retention scores, the number of correct responses in the respective delayed cued recall test was divided by the number of correct responses in the immediate cued recall test in which the learning criterion was reached for each participant. For the recognition performance score, false-alarm rates were subtracted from hit rates to obtain a response-bias-corrected recognition measure per participant¹⁸.

Cued recall and recognition differences between conditions were analyzed by means of one-factorial between-participants ANOVA. These were followed up by multiple planned contrasts to infer the significance of the pairwise differences between the respective conditions of interest²⁵.

We obtained MPT parameter estimates from two different estimation approaches to ensure the robustness of our model-based conclusions. First, using response category frequencies aggregated within experimental conditions, we applied an ML

Measure	Waking rest	Social media	Vocabulary
Immediate recall: Repetitions	1.39 (0.53)	1.48 (0.50)	1.38 (0.60)
Immediate recall: Correct responses	11.59 (2.66)	11.94 (3.56)	10.96 (2.69)
Delayed recall: Correct responses	13.39 (3.11)	13.46 (3.98)	11.81 (3.37)
Recall retention	1.16 (0.15)	1.14 (0.18)	1.08 (0.17)
Hit rate	0.95 (0.08)	0.94 (0.07)	0.88 (0.11)
False-alarm rate	0.05 (0.06)	0.04 (0.05)	0.04 (0.06)
Recognition performance	0.90 (0.10)	0.91 (0.09)	0.83 (0.13)

Table 1. Mean (*SD*) cued recall and recognition performances in Session 1 of Experiment 1. A total of 20 word pairs was presented to participants during the original learning phase. The learning phase and the immediate cued recall were presented between one and three times to participants. Recall retention = correct responses in the delayed cued recall / correct responses in the immediate cued recall. Recognition performance = hit rate – false-alarm rate.

estimation approach as implemented in the R package MPTinR²⁶. Second, for the individual response category frequencies, we applied the Bayesian hierarchical latent-trait estimation approach²⁷ as implemented in the R package TreeBUGS²⁸. Whereas the first approach rests on the assumption of identically and independently distributed (i.i.d.) observations, the latter takes into account the potential heterogeneity of participants and includes parameter correlations²⁸.

For the aggregated data, the model was estimated simultaneously for all three experimental conditions. By applying equality constraints between conditions, a model version with $df > 0$ was obtained that allowed for testing the model via the G^2 goodness-of-fit statistic. For the individual data, the model was estimated separately for each experimental condition. Convergence of the MCMC sampler was confirmed according to the potential scale reduction factor $\hat{R} < 1.05$ ²⁹. Model fit was evaluated with respect to the posterior-predictive p -values obtained from comparing the fit statistics T_1 and T_2 for the observed and posterior-predicted data²⁷. p_{T_1} and p_{T_2} values > 0.05 are considered to reflect satisfactory model fit²⁸.

For the aggregated data, equality constraints were imposed on the respective parameters of interest, and the significance of parameter differences between experimental conditions was inferred from the reduction in model fit, that is, ΔG^2 in the case of two-sided research questions, and $z = \sqrt{\Delta G^2(1)}$ the case of one-sided hypotheses. For the individual data, posterior distributions of parameter differences were used to infer the reliability of parameter differences between conditions¹⁶. More specifically, for our preregistered one-tailed hypotheses, so-called Bayesian p -values were calculated, that is, the relative amount of the posterior distribution below zero. For two-tailed research questions, 95% Bayesian credibility intervals (BCI) were used as a reliability criterion. Given that aggregated and individual MPT results converged in most cases (as has previously been observed for a range of different MPT models³⁰ and is also expected for models of the type relevant here³¹), Bayesian p -values and 95% BCI are only reported when inferences differed from those obtained for the aggregated data. The remaining values are provided in the Supplementary Material.

Results

Manipulation check

The mean number of Instagram posts that participants in the social media condition reported to have engaged in was $M = 23.28$ ($SD = 13.15$, $range = 6-70$). Only one participant reported to have engaged in any Icelandic posts. The mean number of correct responses in the interpolated cued recall in the vocabulary condition was $M = 12.43$ ($SD = 4.70$).

Cued recall and recognition

An inspection of cued recall and recognition measures in the full sample revealed some rather severe outliers in both Sessions 1 and 2. To avoid biased results while not overly compromising the statistical power of our hypothesis tests, we decided to deviate from our preregistered analysis plan by applying a conservative outlier criterion and excluding extreme values from both sessions separately. We excluded participants whose cued recall retention or recognition performance score was more than three times the median absolute distance (MAD) away from the respective grand median³². This approach resulted in sample sizes of $N_1 = 154$ for Session 1 ($n = 51$ in the waking rest condition, $n = 50$ in the social media condition, $n = 53$ in the vocabulary condition) and $N_2 = 141$ for Session 2 ($n = 49$ in the waking rest and social media conditions, $n = 43$ in the vocabulary condition). The resulting descriptive statistics for the cued recall and recognition measures in Session 1 are provided in Table 1.

The number of repetitions of the immediate cued recall necessary to reach the learning criterion did not differ significantly between conditions, $F(2, 151) = 0.52$, $p = 0.593$, $\eta^2 = 0.01$. The same was true for the number of correct responses in the immediate cued recall in which the learning criterion was reached, $F(2, 151) = 1.42$, $p = 0.245$, $\eta^2 = 0.02$.

We hypothesized that cued recall retention within Session 1 would be higher in the waking rest condition than in the social

Parameter	Waking rest	Social media	Vocabulary
	Aggregated data		
s	0.91 [0.90, 0.93]	0.92 [0.91, 0.94]	0.87 [0.84, 0.89]
r_1	0.98 [0.97, 0.99]	0.98 [0.97, 0.99]	0.96 [0.94, 0.98]
g	0.46 [0.36, 0.55]	0.38 [0.28, 0.48]	0.27 [0.20, 0.33]
r_2	0.73 [0.70, 0.76]	0.73 [0.70, 0.76]	0.68 [0.65, 0.71]
	Individual data		
s	0.94 [0.91, 0.97]	0.94 [0.91, 0.97]	0.89 [0.85, 0.93]
r_1	0.99 [0.97, 1.00]	0.98 [0.97, 1.00]	0.99 [0.97, 1.00]
g	0.57 [0.37, 0.80]	0.36 [0.20, 0.52]	0.21 [0.10, 0.32]
r_2	0.74 [0.69, 0.78]	0.75 [0.68, 0.81]	0.69 [0.63, 0.74]

Table 2. Storage-retrieval multinomial processing tree (MPT) parameter estimates [95% CI] in Session 1 of Experiment 1. s = probability of successful target storage, r_1 = probability of successful target retrieval during recognition, g = probability of guessing 'old' during recognition, r_2 = probability of successful target retrieval during recall. For the aggregated data, the model was fitted using ML estimation in the R package MPTinR²⁶ (95% confidence intervals in brackets), and parameter r_1 was set equal between the waking rest and the social media condition to allow for a model fit evaluation. For the individual data, the model was fitted using Bayesian hierarchical estimation in the R package TreeBUGS²⁸ (95% Bayesian credibility intervals in brackets).

media and vocabulary conditions, respectively. Overall, there was a significant effect of our manipulation on the number of correct responses in the first delayed cued recall, $F(2, 151) = 3.71, p = 0.027, \eta^2 = 0.05$. More importantly, in line with our hypotheses, cued recall retention did differ significantly between conditions as well, $F(2, 151) = 3.48, p = 0.033, \eta^2 = 0.04$. We performed planned contrasts to further evaluate our hypotheses. As expected, cued recall retention was significantly higher in the waking rest than in the vocabulary condition, $t(151) = 2.55, p = 0.006$, Cohen's $d = 0.50$. However, this was not the case when the waking rest condition was compared against the social media condition, $t(151) = 0.71, p = 0.238$, Cohen's $d = 0.14$.

We also hypothesized that recognition performance after the 8-min retention interval in Session 1 would be higher in the waking rest condition than in the social media and the vocabulary conditions, respectively. There again was an overall effect of our manipulation, $F(2, 151) = 7.41, p = 0.001, \eta^2 = 0.09$. Mirroring our findings for cued recall retention, planned contrasts revealed that recognition performances were significantly lower in the vocabulary than in the waking rest condition, $t(151) = 3.15, p = 0.001$, Cohen's $d = 0.62$, whereas recognition performances in the social media condition did not differ significantly from those in the waking rest condition, $t(151) = -0.33, p = 0.627$, Cohen's $d = -0.06$.

Descriptive statistics for Session 2 are provided in Supplementary Table S1. Overall, no significant effects emerged between the waking rest and the social media condition, and the differences between the waking rest and the vocabulary condition were substantially reduced. Indeed, only the effect on cued recall retention remained significant, $t(138) = 1.88, p = 0.031$, Cohen's $d = 0.39$.

We confirmed the robustness of our main conclusions concerning cued recall and recognition measures in a sensitivity analysis (see Supplementary Tables S3 and S5).

Storage-retrieval MPT model probabilities

Our preregistered MPT model specification included an equality constraint on guessing parameter g across conditions for the aggregated data. However, this model version did not fit the data, neither for Session 1, $G^2(2) = 11.34, p = 0.003$, nor for Session 2, $G^2(2) = 12.11, p = 0.002$. Instead, an inspection of parameter estimates resulting from a saturated model version indicated that recognition retrieval parameters r_1 might be similar enough between conditions to allow for an equality constraint, especially between the waking rest and social media conditions. Indeed, such a model version fit the data well both for Session 1, $G^2(1) = 0.42, p = 0.515$, and for Session 2, $G^2(1) = 0.15, p = 0.703$. Thus, we used this model version for further MPT analyses. For the individual data, good convergence was observed for all parameters in all three conditions, all $\hat{R} < 1.05$, and the model fit the data well in both sessions.

MPT parameter estimates for Session 1 are provided in Table 2. Estimates largely aligned between both estimation approaches for the aggregated and individual data. Based on the assumption of diversion retroactive interference, we expected MPT storage probabilities s in Session 1 to be higher in the waking rest than in the social media and vocabulary conditions, respectively. We also hypothesized that storage probabilities would not differ significantly between the social media and the vocabulary conditions. Contrary to our hypothesis, there even was a slight descriptive tendency of higher storage probabilities in the social media than in the waking rest condition, $z = 0.73, p = 0.767$. In contrast, storage probabilities were significantly higher in the waking rest compared to the vocabulary condition, $z = 2.99, p = 0.001$. Against our expectations, there was also

a significant difference in storage probabilities between the social media and the vocabulary condition, $\Delta G^2(1) = 13.40$, $p < 0.001$.

Based on the assumption of similarity retroactive interference, we hypothesized that cued recall retrieval probabilities r_2 in Session 1 would be higher in the waking rest than in the social media and the vocabulary conditions, respectively, and also in the social media compared to the vocabulary condition. Against our expectations, cued recall retrieval probabilities in Session 1 did not differ significantly between the waking rest and the social media condition, $z = 0.20$, $p = 0.421$. For the remaining pairwise comparisons, the results were rather mixed: For the individual data, cued recall retrieval probabilities were not reliably higher in the waking rest than in the vocabulary condition, Bayesian $p = 0.096$, or in the social media than in the vocabulary condition, Bayesian $p = 0.082$. In contrast, for the aggregated data, these comparisons yielded significantly higher cued recall retrieval probabilities in the waking rest than in the vocabulary condition, $z = 2.21$, $p = 0.014$, and in the social media than in the vocabulary condition, $z = 2.03$, $p = 0.021$.

With respect to the remaining parameters, recognition retrieval probabilities r_1 were estimated to be very close to 1 in all three conditions. For guessing probabilities g , we observed a descriptive reduction of the probability to guess ‘old’ from the waking rest to the vocabulary condition. Indeed, guessing probabilities did differ significantly between the waking rest and the vocabulary condition, $\Delta G^2(1) = 10.65$, $p = 0.001$. However, this was not the case for the comparisons between the waking rest and social media conditions, $\Delta G^2(1) = 1.16$, $p = 0.282$, or the social media and vocabulary conditions, $\Delta G^2(1) = 3.80$, $p = 0.051$.

MPT parameter estimates for Session 2 are provided in Supplementary Table S2. Whereas storage probabilities s did not differ significantly between conditions anymore, the patterns for parameters r_1 , r_2 and g were very similar to those from Session 1.

We confirmed the robustness of our MPT results in a sensitivity analysis (see Supplementary Tables S4 and S6).

Experiment 2

Methods

Our methodological approach in Experiment 2 largely followed that of Experiment 1, except for a reversal of the recognition-then-cued-recall test procedure and some other deviations and extensions that are detailed below.

Participants

Our sample size rationale was the same as in Experiment 1, that is, we aimed at a statistical power of $1 - \beta = 80\%$ to detect a medium effect size of Cohen’s $f = 0.25$ with $\alpha = 5\%$. However, after observing some rather severe outliers with respect to cued recall retention and recognition performance scores in Experiment 1, we oversampled by 10% to account for the necessity to exclude outliers. Thus, the required sample size was $N = 177$ participants (i.e., $n = 59$ per condition). We preregistered the same exclusion criterion we already used in Experiment 1, that is, participants whose cued recall retention or recognition performance score was more than three times the MAD away from the respective grand median were excluded³².

Participants were again recruited at the University of Mannheim. The same eligibility requirements applied as in Experiment 1, and the same 35% learning criterion was used in the immediate cued recall. Participants could choose between course credit and a financial compensation of 10€ for an estimated net study duration of 45 minutes within one single session.

We collected data from 177 participants. After excluding participants who failed to reach the 35% learning criterion in the immediate cued recall or were identified as outliers according to their recall retention or recognition performance score, the final sample size was $N = 157$ ($n = 53$ in the waking rest condition, $n = 52$ in the social media and vocabulary conditions). Mean age was 22.17 years ($SD = 2.98$, $range = 18-31$). 117 participants (74.52%) indicated to be female, 37 participants (23.57%) indicated to be male, one participant identified their gender as non-binary, and one participant as diverse. One participant refrained from providing their gender identity. 144 participants (91.72%) indicated German to be their first language. As in Experiment 1, all participants confirmed to be fluent in German. 149 participants (94.90%) were studying at the time of participation, and 79 participants (50.32%) were enrolled in a psychology program.

The study was again conducted in accordance with the Declaration of Helsinki (2013). As the study did not involve deception or other ethically relevant elements, formal approval from the ethics committee was not necessary according the regulations of the ethics committee of the University of Mannheim. Informed consent was obtained from all participants.

Procedure

In contrast to Experiment 1, we decided to focus on short-term effects of post-encoding waking rest within a single session. Thus, our procedure in Experiment 2 was the same as in Session 1 of Experiment 1. The delayed testing procedure was reversed, that is, instead of a recognition-then-cued-recall procedure as in Experiment 1, we used a cued-recall-then-recognition procedure. To further increase the sensory input during the 8-min retention interval in the social media condition, participants were asked to bring headphones with them so that they could use Instagram with tone.

Measure	Waking rest	Social media	Vocabulary
Immediate recall: Repetitions	1.49 (0.58)	1.52 (0.58)	1.40 (0.63)
Immediate recall: Correct responses	11.42 (2.95)	12.00 (2.90)	11.62 (2.92)
Delayed recall: Correct responses	11.62 (3.03)	12.17 (2.85)	11.25 (3.39)
Recall retention	1.02 (0.10)	1.02 (0.13)	0.96 (0.11)
Hit rate	0.95 (0.05)	0.94 (0.07)	0.91 (0.08)
False-alarm rate	0.03 (0.06)	0.02 (0.04)	0.03 (0.04)
Recognition performance	0.92 (0.07)	0.92 (0.09)	0.88 (0.09)

Table 3. Mean (*SD*) cued recall and recognition performances in Experiment 2. A total of 20 word pairs was presented to participants during the original learning phase. The learning phase and the immediate cued recall were presented between one and three times to participants. Recall retention = correct responses in the delayed cued recall / correct responses in the immediate cued recall. Recognition performance = hit rate – false-alarm rate.

We used the German translation of the High Sensitive Person Scale (HSPS-G)^{33,34} and the negative emotionality subscale of the German version of the Big Five Inventory 2 (BFI-2)³⁵ as part of a post-experimental questionnaire to assess sensory processing sensitivity and neuroticism as potential moderators of the waking rest effect³⁶. However, these covariates were not of interest in our present research and will be used in separate analyses not reported here.

Results

Manipulation check

The mean number of Instagram posts that participants in the social media condition reported to have engaged in was $M = 24.13$ ($SD = 21.39$, $range = 6-150$). Two participants reported to have engaged in any Icelandic posts. The mean number of correct responses in the interpolated cued recall in the vocabulary condition was $M = 13.58$ ($SD = 4.14$).

Cued recall and recognition

Descriptive statistics for cued recall and recognition measures are provided in Table 3. The number of repetitions of the immediate cued recall did not differ significantly between conditions, $F(2, 154) = 0.53$, $p = 0.591$, $\eta^2 = 0.01$. The same was true for the number of correct responses in the respective last immediate cued recall, $F(2, 154) = 0.54$, $p = 0.584$, $\eta^2 = 0.01$, and in the delayed cued recall, $F(2, 154) = 1.17$, $p = 0.313$, $\eta^2 = 0.01$.

We hypothesized that cued recall retention would be higher in the waking rest condition than in the social media and the vocabulary conditions, respectively. In line with this, an ANOVA revealed a significant overall effect of our manipulation, $F(2, 154) = 4.98$, $p = 0.008$, $\eta^2 = 0.06$. Planned contrasts showed that cued recall retention was significantly higher in the waking rest condition compared to the vocabulary condition, $t(154) = 2.70$, $p = 0.004$, Cohen's $d = 0.53$, but not compared to the social media condition, $t(154) = -0.09$, $p = 0.534$, Cohen's $d = -0.02$.

We also hypothesized that recognition performance would be higher in the waking rest condition than in the social media and the vocabulary conditions, respectively. Again, an ANOVA revealed a significant effect of our manipulation, $F(2, 154) = 3.61$, $p = 0.029$, $\eta^2 = 0.04$. In line with the pattern for cued recall retention, this overall effect could be attributed to significantly lower recognition performances in the vocabulary than in the waking rest condition, $t(154) = 2.23$, $p = 0.014$, Cohen's $d = 0.43$, whereas recognition performances in the social media condition did not differ significantly from those in the waking rest condition, $t(154) = -0.20$, $p = 0.579$, Cohen's $d = -0.04$.

As in Experiment 1, we confirmed the robustness of our main conclusions concerning cued recall and recognition measures in a sensitivity analysis (see Supplementary Table S7).

Storage-retrieval MPT model probabilities

MPT parameter estimates are provided in Table 4. Again, estimates largely aligned between both estimation approaches. For the aggregated data, the same baseline model that we already used in Experiment 1 fit the data well, $G^2(1) = 1.65$, $p = 0.199$. For the individual data, good convergence was observed for all parameters in all three conditions, all $\hat{R} < 1.05$, and the model fit the data well.

We hypothesized that MPT storage probabilities s would be higher in the waking rest than in the social media and vocabulary conditions, respectively. As in Experiment 1, there instead was a descriptive tendency of higher storage probabilities in the social media compared to the waking rest condition, $z = 0.17$, $p = 0.568$. In contrast, storage probabilities in the vocabulary condition were significantly smaller than in the waking rest condition, $z = 2.25$, $p = 0.012$. In line with these two findings, storage probabilities in the vocabulary condition were also significantly smaller than in the social media condition, $\Delta G^2(1) = 5.83$, $p = 0.016$.

Parameter	Waking rest	Social media	Vocabulary
	Aggregated data		
s	0.92 [0.90, 0.93]	0.92 [0.90, 0.94]	0.89 [0.87, 0.91]
r_1	1.00 [1.00, 1.00]	1.00 [1.00, 1.00]	0.99 [0.99, 1.00]
g	0.41 [0.31, 0.51]	0.27 [0.18, 0.36]	0.27 [0.20, 0.35]
r_2	0.63 [0.60, 0.67]	0.66 [0.63, 0.69]	0.63 [0.60, 0.67]
	Individual data		
s	0.92 [0.90, 0.94]	0.93 [0.91, 0.96]	0.89 [0.87, 0.92]
r_1	1.00 [0.99, 1.00]	1.00 [0.99, 1.00]	0.99 [0.98, 1.00]
g	0.30 [0.12, 0.49]	0.29 [0.13, 0.47]	0.27 [0.14, 0.40]
r_2	0.64 [0.59, 0.68]	0.66 [0.63, 0.70]	0.64 [0.59, 0.69]

Table 4. Storage-Retrieval multinomial processing tree (MPT) parameter estimates [95% CI] in Experiment 2. Parameter s = probability of successfully storing a target word in memory, r_1 = probability of successfully retrieving a target word during recognition, g = probability of guessing 'old' during recognition, r_2 = probability of successfully retrieving a target word during cued recall. For the aggregated data, the model was fitted using ML estimation in the R package MPTinR²⁶ (95% confidence intervals in brackets), and parameter r_1 was set equal between the waking rest and the social media condition to allow for a model fit evaluation. For the individual data, the model was fitted using Bayesian hierarchical estimation in the R package TreeBUGS²⁸ (95% Bayesian credibility intervals in brackets).

We expected MPT cued recall retrieval probabilities r_2 to be higher in the waking rest than in the social media and the vocabulary conditions, respectively, and also in the social media compared to the vocabulary condition. As we had observed for parameter s , we found a descriptive tendency of higher cued recall retrieval probabilities in the social media compared to the waking rest condition, $z = 1.25$, $p = 0.895$. Surprisingly, cued recall retrieval probabilities in the vocabulary condition were not significantly smaller than in the waking rest condition either, $z = 0.05$, $p = 0.518$. The same held true for the comparison of the social media and the vocabulary condition, $z = 1.19$, $p = 0.117$.

As for the remaining parameters, recognition retrieval probabilities r_1 were estimated to be very close to 1. With respect to guessing probabilities g , the pattern was more complex: For the individual data, no reliable differences were observed between any of the three conditions, that is, all pairwise 95% BCI overlapped zero. In contrast, for the aggregated data, guessing probabilities were estimated to be significantly higher in the waking rest compared to the social media condition, $\Delta G^2(1) = 4.02$, $p = 0.045$, and also compared to the vocabulary condition, $\Delta G^2(1) = 4.77$, $p = 0.029$. Estimates did not differ significantly between the social media and the vocabulary condition, $\Delta G^2(1) = 0.00$, $p = 0.981$.

As in Experiment 1, we confirmed the robustness of our MPT results in a sensitivity analysis (see Supplementary Table S8).

Discussion

In the present research, we set out to find direct evidence for a role of consolidation processes in the waking rest effect by replicating and extending the social media study reported by Martini et al.⁵.

Across both experiments, we failed to find any memory differences between the waking rest and social media conditions. Thereby, our results add to a growing body of evidence suggesting that certain distractor tasks might be equally beneficial for consolidation as waking rest⁹. Indeed, it has been argued that waking rest might not always minimize cognitive processing, but instead trigger highly active processes such as mentalizing, mind-wandering, and autobiographical thinking³⁷. In contrast, past research has identified relaxation as an important motivational aspect of social media engagement³⁸. Thus, the waking rest versus social media effect might be susceptible to how much participants engage in effortful cognitive processing during waking rest and social media use.

That said, the striking discrepancy between the original finding of a medium to large waking rest versus social media effect and our own null findings across two experiments might also be explained through procedural differences. Word pairs in our study were presented for 12 instead of just 5 seconds during the learning phase, and we applied a 35% learning criterion that was not used in the original study. Thereby, mean numbers of correct responses in the immediate cued recall were substantially increased in our experiments ($M = 10.96 - 12.00$) compared to what was found by Martini et al. ($M = 7.64$ in the waking rest condition, $M = 7.21$ in the social media condition). Interestingly, encoding strength has been suggested to moderate the waking rest effect⁹ such that the effect might be reduced or even eliminated for relatively long presentation times³⁹. Such an explanation might in principle apply to our results. For the time being, replication failures such as ours call into question the use of waking rest interventions in relevant applied settings.

We did find reliable differences in cued recall retention and recognition performance scores between the waking rest and

vocabulary conditions across both experiments. Disregarding our model-based results, such a finding could be easily explained by more traditional accounts of similarity-based retroactive interference¹⁴. However, our storage-retrieval MPT analyses revealed that these differences were driven solely by storage processes. Thereby, we found first direct behavioral evidence for a role of consolidation in the waking rest effect.

Our result pattern with respect to MPT storage probabilities s suggests that some diversion threshold needs to be reached before consolidation processes are inhibited. Apparently, only the very high degree of intentional encoding demands induced in the vocabulary condition was sufficient to interfere with consolidation. This observation contradicts the original key assumption by Dewar et al.³ according to which any interpolated material or task that induces cognitive processing and encoding demands beyond mere waking rest will interfere with consolidation. However, more research is needed to determine whether or not such storage effects can also be observed for low similarity distractor tasks. An ideal candidate to use in such an investigation might be the d2 test of attention⁴⁰, that is, a highly controlled attention and concentration performance test for which negative effects compared to waking rest have recently been demonstrated for some participants^{36,41}. Such a non-verbal task can be assumed to share virtually no similarities with the original learning task while inducing considerable cognitive processing demands.

Our mixed findings in Experiment 1 and clear null-findings in Experiment 2 for cued recall retrieval probabilities r_2 lead to a rather complex conclusion with respect to similarity retroactive interference. Apparently, the similarities between Icelandic-German and unrelated Norwegian-German word pairs were insufficient to result in significant retrieval competition⁴² or indistinctiveness^{43,44} during the delayed memory tests. One reason for this might be that our testing procedure did not involve free recall tests, that is, participants were always presented with some retrieval cue: either the cue word of the respective word pair (cued recall) or even the target word itself (recognition). It seems likely that retrieval differences would have been more pronounced had we used a testing procedure involving free recall. Future research might use alternative storage-retrieval MPT models that involve such free recall tests^{45,46}.

To our surprise, we found mixed evidence for an effect of our manipulation on guessing probabilities g during recognition. Descriptively, the probability of guessing 'old' was highest in the waking rest and lowest in the vocabulary condition across both experiments. Although this pattern was only reliable in Experiment 1, our results at least tentatively suggest that positive effects of post-encoding waking rest on recognition performances might be partially explained by a more balanced guessing style (i.e., g closer to 0.5 in the case of equal numbers of targets and distractors) in the waking rest condition compared to an overly conservative guessing style (i.e., g closer to 0.0) in distractor conditions.

Overall, the MPT result patterns largely aligned between both experiments, suggesting that parameter estimates resulting from this model are rather robust against changes to the test order. However, mean cued recall retention scores above 1 in all three conditions in Experiment 1 suggest that participants benefited from being presented with all target words during the preceding recognition test. In contrast, in Experiment 2, the reversal of the memory testing procedure led to a reduction in cued recall retention scores. Thus, the overall performance level in Experiment 2 might be more trustworthy.

Using storage-retrieval MPT modeling allowed us to directly measure consolidation contributions to memory performance on a behavioral level. We hope that future research on the positive effects of post-encoding waking rest will adopt such a model-based approach. We are optimistic that the field will thereby reach an even more comprehensive and nuanced view of the respective roles of diversion and similarity retroactive interference in the waking rest effect.

Data availability

The datasets generated during the current studies and R scripts necessary to reproduce all reported results are available on the OSF at osf.io/k2gs8/?view_only=f06d32611f1c4808be3ad66fd7d3b973.

References

1. Lamp, A., Cook, M., Soriano Smith, R. N. & Belenky, G. Exercise, nutrition, sleep, and waking rest? *Sleep* **42**, zsz138, DOI: [10.1093/sleep/zsz138](https://doi.org/10.1093/sleep/zsz138) (2019).
2. Wamsley, E. J. Memory consolidation during waking rest. *Trends Cogn. Sci.* **23**, 171–173, DOI: [10.1016/j.tics.2018.12.007](https://doi.org/10.1016/j.tics.2018.12.007) (2019).
3. Dewar, M., Cowan, N. & Della Sala, S. Forgetting due to retroactive interference: A fusion of Müller and Pilzecker's (1900) early insights into everyday forgetting and recent research on anterograde amnesia. *Cortex* **43**, 616–634, DOI: [10.1016/S0010-9452\(08\)70492-1](https://doi.org/10.1016/S0010-9452(08)70492-1) (2007).
4. Humiston, G. B., Tucker, M. A., Summer, T. & Wamsley, E. J. Resting states and memory consolidation: A preregistered replication and meta-analysis. *Sci. Reports* **9**, Article 19345, DOI: [10.1038/s41598-019-56033-6](https://doi.org/10.1038/s41598-019-56033-6) (2019).
5. Martini, M., Heinz, A., Hinterholzer, J., Martini, C. & Sachse, P. Effects of wakeful resting versus social media usage after learning on the retention of new memories. *Appl. Cogn. Psychol.* **34**, 551–558, DOI: [10.1002/acp.3641](https://doi.org/10.1002/acp.3641) (2020).

6. Alber, J., Della Sala, S. & Dewar, M. Minimizing interference with early consolidation boosts 7-day retention in amnesic patients. *Neuropsychology* **28**, 667–675, DOI: [10.1037/neu0000091](https://doi.org/10.1037/neu0000091) (2014).
7. Eccles, D. W., Balk, Y., Gretton, T. W. & Harris, N. “The forgotten session”: Advancing research and practice concerning the psychology of rest in athletes. *J. Appl. Sport Psychol.* **34**, 3–24, DOI: [10.1080/10413200.2020.1756526](https://doi.org/10.1080/10413200.2020.1756526) (2020).
8. Martini, M., Martini, C., Bernegger, C. & Sachse, P. Post-encoding wakeful resting supports the retention of new verbal memories in children aged 13–14 years. *Br. J. Dev. Psychol.* **37**, 199–210, DOI: [10.1111/bjdp.12267](https://doi.org/10.1111/bjdp.12267) (2019).
9. Martini, M. & Sachse, P. Factors modulating the effects of waking rest on memory. *Cogn. Process.* **21**, 149–153, DOI: [10.1007/s10339-019-00942-x](https://doi.org/10.1007/s10339-019-00942-x) (2020).
10. Pittelkow, M.-M. *et al.* The process of replication target selection in psychology: What to consider? *Royal Soc. Open Sci.* **10**, 210586, DOI: [10.1098/rsos.210586](https://doi.org/10.1098/rsos.210586) (2023).
11. Dewar, M., Alber, J., Butler, C., Cowan, N. & Della Sala, S. Brief wakeful resting boosts new memories over the long term. *Psychol. Sci.* **23**, 955–960, DOI: [10.1177/0956797612441220](https://doi.org/10.1177/0956797612441220) (2012).
12. Mercer, T. Wakeful rest alleviates interference-based forgetting. *Memory* **23**, 127–137, DOI: [10.1080/09658211.2013.872279](https://doi.org/10.1080/09658211.2013.872279) (2015).
13. Mednick, S. C., Cai, D. J., Shuman, T., Anagnostaras, S. & Wixted, J. T. An opportunistic theory of cellular and systems consolidation. *Trends Neurosci.* **34**, 504–514, DOI: [10.1016/j.tins.2011.06.003](https://doi.org/10.1016/j.tins.2011.06.003) (2011).
14. McGeoch, J. A. & McDonald, W. T. Meaningful relation and retroactive inhibition. *The Am. J. Psychol.* **43**, 579–588, DOI: [10.2307/1415159](https://doi.org/10.2307/1415159) (1931).
15. Erdfelder, E. *et al.* Multinomial processing tree models: A review of the literature. *Zeitschrift für Psychol. / J. Psychol.* **217**, 108–124, DOI: [10.1027/0044-3409.217.3.108](https://doi.org/10.1027/0044-3409.217.3.108) (2009).
16. Schmidt, O., Erdfelder, E. & Heck, D. W. How to develop, test, and extend multinomial processing tree models: A tutorial. *Psychol. Methods* Advance online publication, DOI: [10.1037/met0000561](https://doi.org/10.1037/met0000561) (2023).
17. Riefer, D. M. & Batchelder, W. H. A multinomial modeling analysis of the recognition-failure paradigm. *Mem. & Cogn.* **23**, 611–630, DOI: [10.3758/BF03197263](https://doi.org/10.3758/BF03197263) (1995).
18. Nadarevic, L. Emotionally enhanced memory for negatively arousing words: Storage or retrieval advantage? *Cogn. & Emot.* **31**, 1557–1570, DOI: [10.1080/02699931.2016.1242477](https://doi.org/10.1080/02699931.2016.1242477) (2017).
19. Moshagen, M. multiTree: A computer program for the analysis of multinomial processing tree models. *Behav. Res. Methods* **42**, 42–54, DOI: [10.3758/BRM.42.1.42](https://doi.org/10.3758/BRM.42.1.42) (2010).
20. Quevedo Pütter, J., Erdfelder, E. & Schnieders, B. A storage-retrieval MPT analysis of the wakeful resting effect, DOI: [10.17605/OSF.IO/ZRJ94](https://doi.org/10.17605/OSF.IO/ZRJ94) (2023).
21. Quevedo Pütter, J. & Erdfelder, E. An adapted storage-retrieval MPT analysis of the wakeful resting effect, DOI: [10.17605/OSF.IO/SMX5Q](https://doi.org/10.17605/OSF.IO/SMX5Q) (2023).
22. Faul, F., Erdfelder, E., Lang, A.-G. & Buchner, A. G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behav. Res. Methods* **39**, 175–191, DOI: [10.3758/BF03193146](https://doi.org/10.3758/BF03193146) (2007).
23. Dimigen, O., Kliegl, R. & Sommer, W. Trans-saccadic parafoveal preview benefits in fluent reading: A study with fixation-related brain potentials. *NeuroImage* **62**, 381–393, DOI: [10.1016/j.neuroimage.2012.04.006](https://doi.org/10.1016/j.neuroimage.2012.04.006) (2012).
24. R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing (2021).
25. Thériault, R. Rempsyc: Convenience functions for psychology. *J. Open Source Softw.* **8**, 5466, DOI: [10.21105/joss.05466](https://doi.org/10.21105/joss.05466) (2023).
26. Singmann, H. & Kellen, D. MPTinR: Analysis of multinomial processing tree models in R. *Behav. Res. Methods* **45**, 560–575, DOI: [10.3758/s13428-012-0259-0](https://doi.org/10.3758/s13428-012-0259-0) (2013).
27. Klauer, K. C. Hierarchical multinomial processing tree models: A latent-trait approach. *Psychometrika* **75**, 70–98, DOI: [10.1007/s11336-009-9141-0](https://doi.org/10.1007/s11336-009-9141-0) (2010).
28. Heck, D. W., Arnold, N. R. & Arnold, D. TreeBUGS: An R package for hierarchical multinomial-processing-tree modeling. *Behav. Res. Methods* **50**, 264–284, DOI: [10.3758/s13428-017-0869-7](https://doi.org/10.3758/s13428-017-0869-7) (2018).
29. Gelman, A. & Rubin, D. B. Inference from iterative simulation using multiple sequences. *Stat. Sci.* **7**, 457–472, DOI: [10.1214/ss/1177011136](https://doi.org/10.1214/ss/1177011136) (1992).

30. Singmann, H. *et al.* Evaluating the robustness of parameter estimates in cognitive models: A meta-analytic review of multinomial processing tree models across the multiverse of estimation methods. *Psychol. Bull.* Advance online publication, DOI: [10.1037/bul0000434](https://doi.org/10.1037/bul0000434) (2024).
31. Erdfelder, E., Quevedo Pütter, J. & Schnuerch, M. On aggregation invariance of multinomial processing tree models. *Behav. Res. Methods* (in press).
32. Leys, C., Ley, C., Klein, O., Bernard, P. & Licata, L. Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median. *J. Exp. Soc. Psychol.* **49**, 764–766, DOI: [10.1016/j.jesp.2013.03.013](https://doi.org/10.1016/j.jesp.2013.03.013) (2013).
33. Aron, E. N. & Aron, A. Sensory-processing sensitivity and its relation to introversion and emotionality. *J. Pers. Soc. Psychol.* **73**, 345–368, DOI: [10.1037/0022-3514.73.2.345](https://doi.org/10.1037/0022-3514.73.2.345) (1997).
34. Konrad, S. & Herzberg, P. Y. Psychometric properties and validation of a German High Sensitive Person Scale (HSPS-G). *Eur. J. Psychol. Assess.* **35**, 364–378, DOI: [10.1027/1015-5759/a000411](https://doi.org/10.1027/1015-5759/a000411) (2019).
35. Danner, D. *et al.* Das Big Five Inventar 2. *Diagnostica* **65**, 121–132, DOI: [10.1026/0012-1924/a000218](https://doi.org/10.1026/0012-1924/a000218) (2019).
36. Marhenke, R., Acevedo, B., Sachse, P. & Martini, M. Individual differences in sensory processing sensitivity amplify effects of post-learning activity for better and for worse. *Sci. Reports* **13**, 4451, DOI: [10.1038/s41598-023-31192-9](https://doi.org/10.1038/s41598-023-31192-9) (2023).
37. Varma, S. *et al.* Non-interfering effects of active post-encoding tasks on episodic memory consolidation in humans. *Front. Behav. Neurosci.* **11**, Article 54, DOI: [10.3389/fnbeh.2017.00054](https://doi.org/10.3389/fnbeh.2017.00054) (2017).
38. Whiting, A. & Williams, D. Why people use social media: A uses and gratifications approach. *Qual. Mark. Res.* **16**, 362–369, DOI: [10.1108/QMR-06-2013-0041](https://doi.org/10.1108/QMR-06-2013-0041) (2013).
39. Fatania, J. & Mercer, T. Nonspecific retroactive interference in children and adults. *Adv. Cogn. Psychol.* **13**, 314–322, DOI: [10.5709/acp-0231-6](https://doi.org/10.5709/acp-0231-6) (2017).
40. Brickenkamp, R. Test d2 - Aufmerksamkeits-Belastungs-Test (Hogrefe, 2002).
41. Martini, M., Marhenke, R., Martini, C., Rossi, S. & Sachse, P. Individual differences in working memory capacity moderate effects of post-learning activity on memory consolidation over the long term. *Sci. Reports* **10**, 17976, DOI: [10.1038/s41598-020-74760-z](https://doi.org/10.1038/s41598-020-74760-z) (2020).
42. Anderson, M. C. Rethinking interference theory: Executive control and the mechanisms of forgetting. *J. Mem. Lang.* **49**, 415–445, DOI: [10.1016/j.jml.2003.08.006](https://doi.org/10.1016/j.jml.2003.08.006) (2003).
43. Brown, G. D. A., Neath, I. & Chater, N. A temporal ratio model of memory. *Psychol. Rev.* **114**, 539–576, DOI: [10.1037/0033-295X.114.3.539](https://doi.org/10.1037/0033-295X.114.3.539) (2007).
44. Ecker, U. K. H., Brown, G. D. A. & Lewandowsky, S. Memory without consolidation: Temporal distinctiveness explains retroactive interference. *Cogn. Sci.* **39**, 1570–1593, DOI: [10.1111/cogs.12214](https://doi.org/10.1111/cogs.12214) (2015).
45. Batchelder, W. H. & Riefer, D. M. Separation of storage and retrieval factors in free recall of clusterable pairs. *Psychol. Rev.* **87**, 375–397, DOI: [10.1037/0033-295X.87.4.375](https://doi.org/10.1037/0033-295X.87.4.375) (1980).
46. Küpper-Tetzel, C. E. & Erdfelder, E. Encoding, maintenance, and retrieval processes in the lag effect: A multinomial processing tree analysis. *Memory* **20**, 37–47, DOI: [10.1080/09658211.2011.631550](https://doi.org/10.1080/09658211.2011.631550) (2012).

Acknowledgements

This research was funded by the Deutsche Forschungsgemeinschaft (DFG), grant 2277, Research Training Group Statistical Modeling in Psychology (SMiP). We would like to thank Bastian Schnieders and Tracy Hoang for their help with data collection.

Author contributions statement

JQP: Conceptualization (lead), Methodology, Software, Investigation, Data Curation, Writing - Original Draft, Writing - Review & Editing, Validation. EE: Conceptualization (supporting), Resources, Writing - Review & Editing, Supervision.

Additional information

The authors declare no competing interests.

All materials necessary to replicate the reported studies are available on the OSF at osf.io/k2gs8/?view_only=f06d32611f1c4808be3ad66fd7d3b973.

Supplementary Information

accompanying the manuscript

*Waking rest during retention facilitates memory consolidation, but so does social media use:
A model-based storage-retrieval analysis*

Main cued recall and recognition analysis of Session 2 in Experiment 1

Our expectations with respect to cued recall and recognition measures in Session 2 of Experiment 1 were the same as for Session 1. We observed no significant overall effect of our manipulation on the number of correct responses in the second delayed cued recall, $F(2, 138) = 1.96, p = 0.145, \eta^2 = 0.03$, or on the cued recall retention across the 24-hr interval between sessions, $F(2, 138) = 1.79, p = 0.171, \eta^2 = 0.03$. In line with this, planned contrasts revealed no significant difference between the waking rest and the social media condition, $t(138) = 0.78, p = 0.220$, Cohen's $d = 0.16$. Importantly, however, cued recall retention in the vocabulary condition was significantly lower than in the waking rest condition, $t(138) = 1.88, p = 0.031$, Cohen's $d = 0.39$.

Recognition performances in Session 2 were overall not significantly affected by our manipulation, $F(2, 138) = 1.88, p = 0.157, \eta^2 = 0.03$. Accordingly, planned contrasts revealed no significant differences between the waking rest and the social media condition, $t(138) = -0.46, p = 0.676$, Cohen's $d = -0.09$, or between the waking rest and the vocabulary condition, $t(138) = 1.43, p = 0.078$, Cohen's $d = 0.30$.

Measure	Waking rest	Social media	Vocabulary
Delayed recall: Correct responses	13.29 (3.11)	13.35 (4.08)	12.02 (3.45)
Recall retention	1.15 (0.21)	1.12 (0.24)	1.07 (0.19)
Hit rate	0.94 (0.07)	0.93 (0.08)	0.90 (0.09)
False-alarm rate	0.04 (0.05)	0.02 (0.04)	0.03 (0.05)
Recognition performance	0.89 (0.09)	0.90 (0.08)	0.87 (0.10)

Supplementary Table S1. Mean (*SD*) cued recall and recognition performances in the main analysis of Session 2 in Experiment 1. Recall retention = correct responses in the delayed cued recall / correct responses in the immediate cued recall. Recognition performance = hit rate – false-alarm rate.

Main Bayesian hierarchical MPT analysis of Session 1 in Experiment 1

The model fit the individual data from Session 1 well ($p_{T1} = 0.502$, $p_{T2} = 0.549$ in the waking rest condition, $p_{T1} = 0.509$, $p_{T2} = 0.491$ in the social media condition, $p_{T1} = 0.502$, $p_{T2} = 0.521$ in the vocabulary condition).

Storage probabilities s in Session 1 were not reliably higher in the waking rest than in the social media condition, Bayesian $p = 0.530$. In contrast, storage probabilities were reliably higher in the waking rest than in the vocabulary condition, Bayesian $p = 0.029$. There was also a reliable difference in storage probabilities between the social media and the vocabulary condition, 95% BCI = [0.00, 0.10].

Cued recall retrieval probabilities r_2 in Session 1 were not reliably higher in the waking rest than in the social media condition, Bayesian $p = 0.631$.

Guessing probabilities g in Session 1 differed reliably between the waking rest and the vocabulary condition, 95% BCI = [0.14, 0.62], but not between the waking rest and the social media condition, 95% BCI = [-0.05, 0.49], or the social media and the vocabulary condition, 95% BCI = [-0.04, 0.34].

Main MPT analysis of Session 2 in Experiment 1

In Session 2 of Experiment 1, storage probabilities s did not differ reliably between any conditions, $z = 0.51$, $p = 0.695$, Bayesian $p = 0.450$ (waking rest versus social media), $z = 1.15$, $p = 0.125$, Bayesian $p = 0.190$ (waking rest versus vocabulary), $\Delta G^2(1) = 2.68$, $p = 0.101$, 95% BCI = [-0.03, 0.06] (social media versus vocabulary). In contrast, the pattern for cued recall retrieval probabilities r_2 was the same as in Session 1, $z = 0.09$, $p = 0.465$, Bayesian $p = 0.624$ (waking rest versus social media), $z = 2.30$, $p = 0.011$, Bayesian $p = 0.081$ (waking rest versus vocabulary), $z = 2.26$, $p = 0.012$, Bayesian $p = 0.072$ (social media versus vocabulary). Again, recognition retrieval parameters r_1 were estimated to be very close to 1. The pattern for guessing probabilities g was also very similar to that in Session 1, $\Delta G^2(1) = 5.67$, $p = 0.017$, 95% BCI = [-0.05, 0.40] (waking rest versus social media), $\Delta G^2(1) = 11.34$, $p = 0.001$, 95% BCI = [0.07, 0.42] (waking rest versus vocabulary), $\Delta G^2(1) = 0.74$, $p = 0.391$, 95% BCI = [-0.14, 0.29] (social media versus vocabulary).

Parameter	Waking rest	Social media	Vocabulary
Aggregated data			
s	0.91 [0.89, 0.93]	0.91 [0.89, 0.93]	0.89 [0.86, 0.91]
r_1	0.99 [0.98, 1.00]	0.99 [0.98, 1.00]	0.98 [0.96, 0.99]
g	0.41 [0.32, 0.50]	0.25 [0.17, 0.34]	0.20 [0.13, 0.28]
r_2	0.73 [0.70, 0.77]	0.73 [0.70, 0.76]	0.68 [0.64, 0.71]
Individual data			
s	0.93 [0.89, 0.96]	0.92 [0.90, 0.95]	0.91 [0.87, 0.94]
r_1	0.99 [0.97, 1.00]	0.99 [0.98, 1.00]	0.98 [0.96, 0.99]
g	0.39 [0.26, 0.52]	0.21 [0.05, 0.41]	0.14 [0.04, 0.28]
r_2	0.74 [0.70, 0.79]	0.76 [0.69, 0.82]	0.69 [0.63, 0.75]

Supplementary Table S2. Storage-retrieval multinomial processing tree (MPT) parameter estimates (95% CI) in the main analysis of Session 2 in Experiment 1. s = probability of successful target storage, r_1 = probability of successful target retrieval during recognition, g = probability of guessing 'old' during recognition, r_2 = probability of successful target retrieval during recall. For the aggregated data, the model was fitted using ML estimation (95% confidence intervals in brackets), and parameter r_1 was set equal between the waking rest and the social media condition to allow for a model fit evaluation. For the individual data, the model was fitted using Bayesian hierarchical estimation (95% Bayesian credibility intervals in brackets).

Sensitivity analysis of Sessions 1 and 2 in Experiment 1

We conducted a sensitivity analysis to check the robustness of our main conclusions from Experiment 1. For Session 1, the respective analyses were conducted without the data from those participants who reported in the post-experimental questionnaire to have consciously rehearsed any of the Icelandic-German vocabulary during the 8-min retention interval ($n = 55$). For Session 2, they were conducted without the data from those participants who reported (a) 4 hours of sleep or less during the night between sessions ($n = 2$), (b) alcohol consumption between sessions ($n = 31$), (c) conscious rehearsal of the Icelandic-German vocabulary during the 24-hr interval between sessions ($n = 55$), (d) correct assumptions about our hypotheses ($n = 1$), or (e) a lack of understanding of the study instructions ($n = 0$). Overall, data from $N_1 = 99$ participants was included in the sensitivity analysis for Session 1 ($n = 30$ in the waking rest condition, $n = 38$ in the social media condition, $n = 31$ in the vocabulary condition), and data from $N_2 = 65$ participants in the sensitivity analysis for Session 2 ($n = 29$ in the waking rest condition, $n = 23$ in the social media condition, $n = 13$ in the vocabulary condition).

Descriptive statistics for cued recall and recognition performances are provided in Table S3 for Session 1 and in Table S5 for Session 2, MPT parameter estimates from both estimation approaches are provided in Table S4 for Session 1 and in Table S6 for Session 2. With respect to our dependent variables of main interest (i.e., cued recall retention, recognition performance, storage probability s , cued recall retrieval probability r_2), the overall data patterns largely aligned with those obtained from the main analyses. This tentatively indicates that our main conclusions are robust against conscious rehearsal, sleep deprivation, alcohol consumption, and participants' awareness of our hypotheses. However, given the considerably smaller sample sizes in the sensitivity compared to the main analyses, we deemed further significance testing uncalled-for. Thus, any data patterns observed in the sensitivity analysis should be treated with caution.

Supplementary Information

Measure	Waking rest	Social media	Vocabulary
Immediate recall: Repetitions	1.40 (0.56)	1.39 (0.50)	1.32 (0.54)
Immediate recall: Correct responses	11.83 (2.67)	11.42 (3.39)	10.84 (2.45)
Delayed recall: Correct responses	13.70 (3.10)	12.82 (3.89)	11.71 (3.49)
Recall retention	1.17 (0.13)	1.13 (0.18)	1.07 (0.17)
Hit rate	0.95 (0.07)	0.94 (0.06)	0.86 (0.12)
False-alarm rate	0.05 (0.07)	0.03 (0.05)	0.03 (0.05)
Recognition performance	0.90 (0.11)	0.91 (0.08)	0.83 (0.13)

Supplementary Table S3. Mean (*SD*) cued recall and recognition performances in the sensitivity analysis of Session 1 in Experiment 1. Recall retention = correct responses in the delayed cued recall / correct responses in the immediate cued recall. Recognition performance = hit rate – false-alarm rate.

Parameter	Waking rest	Social media	Vocabulary
Aggregated data			
s	0.91 [0.89, 0.94]	0.93 [0.90, 0.95]	0.88 [0.85, 0.91]
r_1	0.98 [0.97, 0.99]	0.98 [0.97, 0.99]	0.94 [0.92, 0.97]
g	0.48 [0.35, 0.60]	0.37 [0.26, 0.49]	0.17 [0.10, 0.24]
r_2	0.75 [0.71, 0.79]	0.69 [0.66, 0.73]	0.67 [0.62, 0.71]
Individual data			
s	0.95 [0.90, 0.98]	0.93 [0.91, 0.96]	0.91 [0.86, 0.96]
r_1	0.99 [0.96, 1.00]	0.99 [0.97, 1.00]	0.98 [0.94, 1.00]
g	0.49 [0.25, 0.74]	0.31 [0.13, 0.50]	0.12 [0.03, 0.24]
r_2	0.75 [0.70, 0.81]	0.71 [0.64, 0.78]	0.67 [0.59, 0.75]

Supplementary Table S4. Storage-retrieval multinomial processing tree (MPT) parameter estimates [95% CI] in the sensitivity analysis of Session 1 in Experiment 1. s = probability of successful target storage, r_1 = probability of successful target retrieval during recognition, g = probability of guessing 'old' during recognition, r_2 = probability of successful target retrieval during recall. For the aggregated data, the model was fitted using ML estimation (95% confidence intervals in brackets), and parameter r_1 was set equal between the waking rest and the social media condition to allow for a model fit evaluation. For the individual data, the model was fitted using Bayesian hierarchical estimation (95% Bayesian credibility intervals in brackets).

Supplementary Information

Measure	Waking rest	Social media	Vocabulary
Delayed recall: Correct responses	13.48 (2.89)	11.78 (3.59)	12.38 (3.50)
Recall retention	1.17 (0.22)	1.11 (0.28)	1.09 (0.17)
Hit rate	0.94 (0.06)	0.91 (0.09)	0.87 (0.10)
False-alarm rate	0.04 (0.05)	0.03 (0.04)	0.02 (0.03)
Recognition performance	0.90 (0.08)	0.88 (0.08)	0.85 (0.10)

Supplementary Table S5. Mean (*SD*) cued recall and recognition performances in the sensitivity analysis of Session 2 in Experiment 1. Recall retention = correct responses in the delayed cued recall / correct responses in the immediate cued recall. Recognition performance = hit rate – false-alarm rate.

Parameter	Waking rest	Social media	Vocabulary
Aggregated data			
s	0.92 [0.89, 0.94]	0.89 [0.86, 0.92]	0.87 [0.82, 0.91]
r_1	0.99 [0.98, 1.00]	0.99 [0.98, 1.00]	0.99 [0.97, 1.01]
g	0.37 [0.25, 0.49]	0.24 [0.13, 0.35]	0.11 [0.01, 0.20]
r_2	0.74 [0.70, 0.78]	0.66 [0.61, 0.71]	0.72 [0.65, 0.78]
Individual data			
s	0.93 [0.89, 0.97]	0.90 [0.85, 0.93]	0.88 [0.79, 0.95]
r_1	0.98 [0.96, 1.00]	0.99 [0.97, 1.00]	0.98 [0.95, 1.00]
g	0.36 [0.20, 0.53]	0.22 [0.03, 0.54]	0.12 [0.01, 0.37]
r_2	0.74 [0.68, 0.80]	0.67 [0.58, 0.75]	0.73 [0.59, 0.85]

Supplementary Table S6. Storage-retrieval multinomial processing tree (MPT) parameter estimates [95% CI] in the sensitivity analysis of Session 2 in Experiment 1. s = probability of successful target storage, r_1 = probability of successful target retrieval during recognition, g = probability of guessing 'old' during recognition, r_2 = probability of successful target retrieval during recall. For the aggregated data, the model was fitted using ML estimation (95% confidence intervals in brackets), and parameter r_1 was set equal between the waking rest and the social media condition to allow for a model fit evaluation. For the individual data, the model was fitted using Bayesian hierarchical estimation (95% Bayesian credibility intervals in brackets).

Main Bayesian hierarchical MPT analysis of Experiment 2

The model fit the individual data well ($p_{T1} = 0.501$, $p_{T2} = 0.501$ in the waking rest condition, $p_{T1} = 0.461$, $p_{T2} = 0.533$ in the social media condition, $p_{T1} = 0.496$, $p_{T2} = 0.464$ in the vocabulary condition).

Storage probabilities s were not reliably higher in the waking rest than in the social media condition, Bayesian $p = 0.751$. In contrast, storage probabilities were reliably higher in the waking rest than in the vocabulary condition, Bayesian $p = 0.045$. There was also a reliable difference in storage probabilities between the social media and the vocabulary condition, 95% BCI = [0.00, 0.08].

Cued recall retrieval probabilities r_2 in Session 1 were not reliably higher in the waking rest than in the social media condition, Bayesian $p = 0.809$. The same held true for the comparisons between the waking rest and the vocabulary condition, Bayesian $p = 0.536$, as well as between the social media and the vocabulary condition, Bayesian $p = 0.235$.

Guessing probabilities g did not differ reliably between the waking rest and the social media condition, 95% BCI = [-0.23, 0.26], the waking rest and the vocabulary condition, 95% BCI = [-0.19, 0.26], or the social media and the vocabulary condition, 95% BCI = [-0.18, 0.23].

Sensitivity analysis of Experiment 2

As in Experiment 1, we conducted a sensitivity analysis to check the robustness of our main conclusions. The respective analyses were conducted without the data from those participants who reported in the post-experimental questionnaire (a) to have consciously rehearsed any of the Icelandic-German vocabulary during the 8-min retention interval ($n = 73$), (b) correct assumptions about our hypotheses ($n = 5$), or (c) a lack of understanding of the study instructions ($n = 0$). Overall, data from $N = 83$ participants was included in the sensitivity analysis ($n = 21$ in the waking rest condition, $n = 30$ in the social media condition, $n = 32$ in the vocabulary condition).

Descriptive statistics for cued recall and recognition performances are provided in Table S7, MPT parameter estimates from both estimation approaches in Table S8. With respect to our dependent variables of main interest (i.e., cued recall retention, recognition performance, storage probability s , cued recall retrieval probability r_2), the overall data patterns largely aligned with those obtained from the main analyses. This indicates that our main conclusions are robust against conscious rehearsal and participants' awareness of our hypotheses. Again, sample sizes were considerably smaller than in the main analysis, so any data patterns observed in the sensitivity analysis should be treated with caution.

Supplementary Information

Measure	Waking rest	Social media	Vocabulary
Immediate recall: Repetitions	1.33 (0.58)	1.37 (0.61)	1.34 (0.65)
Immediate recall: Correct responses	10.43 (2.91)	11.40 (2.92)	10.91 (2.79)
Delayed recall: Correct responses	10.57 (2.89)	11.70 (2.96)	10.41 (3.19)
Recall retention	1.02 (0.08)	1.03 (0.11)	0.95 (0.12)
Hit rate	0.95 (0.05)	0.93 (0.08)	0.90 (0.08)
False-alarm rate	0.05 (0.07)	0.01 (0.03)	0.03 (0.05)
Recognition performance	0.90 (0.09)	0.92 (0.09)	0.87 (0.08)

Supplementary Table S7. Mean (*SD*) cued recall and recognition performances in the sensitivity analysis of Experiment 2. Recall retention = correct responses in the delayed cued recall / correct responses in the immediate cued recall. Recognition performance = hit rate – false-alarm rate.

Parameter	Waking rest	Social media	Vocabulary
Aggregated data			
s	0.90 [0.87, 0.93]	0.92 [0.89, 0.94]	0.88 [0.85, 0.90]
r_1	1.00 [1.00, 1.00]	1.00 [1.00, 1.00]	0.99 [0.98, 1.00]
g	0.46 [0.31, 0.61]	0.16 [0.06, 0.26]	0.26 [0.17, 0.35]
r_2	0.59 [0.53, 0.64]	0.64 [0.60, 0.68]	0.59 [0.55, 0.64]
Individual data			
s	0.92 [0.87, 0.96]	0.93 [0.90, 0.96]	0.88 [0.85, 0.91]
r_1	0.99 [0.98, 1.00]	1.00 [0.99, 1.00]	0.99 [0.97, 1.00]
g	0.40 [0.12, 0.69]	0.13 [0.02, 0.30]	0.23 [0.09, 0.39]
r_2	0.58 [0.50, 0.66]	0.64 [0.58, 0.70]	0.60 [0.53, 0.66]

Supplementary Table S8. Storage-retrieval multinomial processing tree (MPT) parameter estimates (95% CI) in the sensitivity analysis of Experiment 2. s = probability of successful target storage, r_1 = probability of successful target retrieval during recognition, g = probability of guessing 'old' during recognition, r_2 = probability of successful target retrieval during cued recall. For the aggregated data, the model was fitted using ML estimation (95% confidence intervals in brackets), and parameter r_2 was set equal between the waking rest and the social media condition to allow for a model fit evaluation. A small positive constant of 0.10 was added to all category frequencies to avoid convergence issues due to categories with zero frequencies. For the individual data, the model was fitted using Bayesian hierarchical estimation (95% Bayesian credibility intervals in brackets).