Title

Enhancing learning with subsecond retrieval attempts AND Practice testing of science text material enhances retention of practiced material, but non-practiced material is unaffected

Author

Reilly, Walter Bernard

Publication Date

2023

Peer reviewed|Thesis/dissertation

Enhancing learning with subsecond retrieval attempts
AND
Practice testing of science text material enhances retention of practiced material, but non-practiced material is unaffected

By

WALTER BERNARD REILLY

DISSERTATION

Submitted in partial satisfaction of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

Psychology

in the

OFFICE OF GRADUATE STUDIES

of the

UNIVERSITY OF CALIFORNIA

DAVIS

Approved:

_____
Charan Ranganath, Chair

_____
Tamara Swaab

_____
Randall O'Reilly

Committee in Charge

2023

ACKNOWLEDGEMENTS

I thank my advisor Charan Ranganath for providing me the opportunity and environment to pursue my intellectual curiosity. You believed in my potential and guided my development as a scientist with compassion. I thank my friend and collaborator James Antony for being the mentor I needed. Your feedback, attentiveness, open yet incisive commentary, and your example, have been instrumental in my completion of this dissertation. I thank Alexander Barnett for your friendship, collaboration, and dedication to science. I thank Andrew Yonelinas for your example of balance in science, teaching, and extracurriculars. I thank the additional members of my dissertation proposal committee, Tamara Swaab, Randall O'Reilly, and Scott Hinze. I thank the Dynamic Memory Lab and its members, past and present. I thank Jordan, Kamin, Zach, and Nichole for your companionship. I thank the Center for Neuroscience and the Department of Psychology, your members, students, and staff. I thank Angela Scully for all that you do for our program. I thank Mark McDaniel and the Memory and Complex Learning Lab for spurring my interest in memory and education and setting the mold of my scientific mind. I thank Toshiya Miyatsu for your mentorship and the confidence you instilled. I thank Sharda Umanath for making my PhD a possibility. You nudged me and then helped along with everything I didn't know I needed. I thank all of my teachers and faculty, going all the way back.

I thank the scholarship benefactors of Saint Louis University High School and Washington University in Saint Louis. I thank the National Science Foundation Graduate Research Fellowship Program.

I thank my friends who have supported my academic endeavors directly and indirectly. I thank Mackenzie Englund for your friendship, support, and stoke.

I thank my Mom and Dad, who gave me every opportunity they could. I thank Myrtie, Louis, April, Owen, Diana, Vernon, Peggy, Wyndel, Pat, Marla, and my grandparents.

I thank Katie, my partner, through this and all the rest.

TABLE OF CONTENTS

LIST OF FIGURES

Chapter 1

Chapter 2

Dissertation Abstract

Memory retrieval is well known to modify retention of not only retrieved material, but also related, non-retrieved material. This dissertation consists of two manuscripts that investigated retrieval phenomena in the learning of foreign language vocabulary and science education material. The testing effect–the retention benefit of practicing retrieval compared to studying–and the pretesting effect–incorrectly guessing a target before learning compared to studying–demonstrate the advantages of retrieval-based learning, but no extant theories have accounted for both of these effects. In Chapter 1, we investigated an error-driven learning account whereby retrieval-based learning serves to "stress test" the memory system, allowing it to learn to better predict a target from a cue, whereas in studying, there is no opportunity for the system to form a prediction. We predicted that inserting a small temporal "gap" between a foreign language word and its English translation should enhance retention when compared to simultaneous, "no gap," presentation. In four experiments ($N = 287$) we consistently observed that "gap" conditions benefitted retention compared to "no gap" conditions, which supports the error-driven learning account. We observed that a gap as short as 600 ms benefitted retention one day later, one minute later, and in a pure list design. In Chapter 2, we investigated the sequalae of retrieval practice for non-practiced educational science text material. Some evidence suggests that retrieval practice of main ideas would lead to greater retention of non-practiced information, whereas other evidence suggests that retrieval practice of peripheral information would lead to impaired retention of non-practiced information, when compared to a control group that did not practice retrieval. In two experiments ($N = 360$) we observed robust testing effects, but we did not observe robust differences in non-practiced material, suggesting that the kind of focal retrieval practice used here has focal effects on science material retention.

Chapter 1: Enhancing learning with subsecond retrieval attempts

Abstract

The testing effect–the retention benefit of practicing retrieval compared to studying–and the pretesting effect–incorrectly guessing a target before learning compared to studying–demonstrate the advantages of retrieval-based learning, but no extant theories have accounted for both of these effects. Error-driven learning is the algorithm underlying the success of neural networks in which robust learning takes place by comparing a predicted pattern to a correct pattern. Here, we argue that retrieval involves the same kind of algorithm that takes place in neural networks. According to the error-driven learning account, retrieval-based learning serves to "stress test" the memory system, allowing it to learn to better predict a target from a cue, whereas in studying, there is no opportunity for the system to form a prediction. An alternative account holds that pretesting effects are due to overt retrieval errors. Here, we present results from four experiments designed to adjudicate between these two accounts. Participants learned Swahili-English translations in two types of conditions. In the "gap" conditions, the cue was presented before the target, creating a gap during which a retrieval attempt could take place. In the "no gap" condition, the cue and the target were presented simultaneously. In three experiments, we found that a gap of only 600 milliseconds enhanced retention of word pairs compared to the no gap condition one day after learning. Participants made very few errors during learning, therefore these results are consistent with the error-driven learning account and are inconsistent with the error-correction account.

Introduction

Educators and students of memory have long sought ways to optimize the allocation of learning time. Retrieval practice, attempting to retrieve a target from a specific cue, has proven to be one of the most potent and time efficient memory enhancement techniques. For example, Carrier and Pashler (1992) showed that retrieval practice enhanced retention more than studying, a phenomenon known as the "testing effect", despite that retrieval practice allows less time with the complete stimulus set. The retention benefits of retrieval practice have borne out across a range of experimental and applied educational settings (Roediger & Karpicke, 2006; Rowland, 2014, Adesope et al., 2017; Agarwal et al., 2021).

Surprisingly, one does not need to attempt to retrieve previously learned material in order to benefit from testing. In studies of the "pretesting effect", participants learn more by guessing a word definition they have never learned than simply reading the word and its definition (e.g., Kornell et al., 2009). The significant learning advantage of retrieval attempts–even without previously learning the target information–is counter-intuitive and prompts the question of why processing an incomplete cue can be superior to encoding a complete study association.

We propose that the testing and pretesting effects may rely on a more general learning mechanism that resembles the kind of error-driven learning (EDL) that occurs in neural network models (Widrow & Hoff, 1960; Rescorla & Wagner, 1972; Rumelhart & McClelland, 1981; for a review, Hoppe et al., 2022). In contrast to the idea that learning involves re-encoding an entire target pattern, many neural networks learn through errors. EDL can be implemented in a number of ways, but in essence, all methods involve comparing a pattern produced by a neural network against a target pattern, and deviations from the target pattern are used to tune the network's underlying representations so that it can more effectively produce a target on subsequent

occasions. In the context of memory, EDL would result from the comparison of a neural pattern produced during retrieval attempts against actual feedback. The intuition is that robust learning occurs when there is error, i.e., when the learning conditions are challenging. Indeed, recent simulations with neural network models of category learning suggest that learning is optimized when the error rate during training is about 15% (Wilson et al., 2019).

How does EDL relate to phenomena like the testing effect and the pretesting effect? If EDL applies to human memory, learners should learn most when learning conditions serve as a "stress test" for memory, identifying and appropriately modifying weak links in a memory representation (Ketz et al., 2013; Liu et al., 2021; Liu & Ranganath, 2021; Mozer et al., 2004; Zheng et al., 2022). Repeated study attempts can be seen as a form of learning without error, as there is no opportunity for the learner to compare their internal representations against the target that is to be learned (Carrier & Pashler, 1992). In contrast, during retrieval practice, only the cue is presented (cue-???), forcing the learner to generate a predicted target. Memory representations are ultimately noisy, so a memory-based response will never fully approximate the retrieval target. As a result, the degree of mismatch between the neural representation of the generated target and the actual target can serve as an error signal, improving the learner's ability to generate the target on subsequent retrieval attempts. Even in a pre-testing paradigm, one can still generate a potential target, and the mismatch between the generated target and the actual target that is to be learned can serve as a potent error signal.

The concept of error-driven learning as incorporated within neural networks inspired a related, but fundamentally distinct, account of the testing effect in Carrier and Pashler (1992). They developed a paradigm to test whether the retention benefits of retrieval practice are caused by mere exposure to correctly retrieved targets. Participants first studied foreign language or

CVC pairs, then in a testing condition a cue was presented for 5 seconds before both the cue and target were presented for five seconds, creating a stimulus-onset asynchrony of targets (SOA = 5 s). In the study condition, cues and targets were presented simultaneously (SOA = 0 s) for a total of ten seconds. In this way, participants could attempt retrieval, but would be exposed to correct answer feedback regardless of whether their retrieval attempt was successful or not. Their results revealed that testing led to greater retention than studying, despite that studied pairs were given more exposure time. Carrier and Pashler (1992) presented an error correction hypothesis that learning can be maximized when one makes an overt error of commission and then has the opportunity to correct the error through feedback. Accordingly, the larger the error committed, the more learning that takes place.

Kornell et al. (2009) and Grimaldi and Karpicke (2012) tested the error correction hypothesis by manipulating the semantic relatedness of paired associates in experiments designed to investigate the mechanisms of the pretesting effect. In the related condition, stimuli consisted of weak associates, such as tide-beach. In the unrelated condition, stimuli were chosen so that pre-experimental associations were unlikely to exist, such as pillow-leaf. In both experiments, a pretesting effect was observed for related pairs but not for unrelated pairs.

Kornell and Grimaldi's results challenge the error correction hypothesis because pretesting should have elicited more error in the unrelated condition than in the related condition. For instance, a participant may have guessed 'wave' in response to the cue 'tide' when the correct target was 'beach', creating a small error, but guessed 'bedroom' in response to the cue 'pillow' when the correct target was 'leaf', resulting in greater error in terms of the semantic distance between the guess and the correct target. Indeed, in a similar pretesting effect study to Grimaldi and Karpicke, Huelser and Metcalfe (2012) found that guesses made during the pretest

phase were equally semantically related to the cue in both the related and unrelated condition, strongly suggesting that guesses in the unrelated condition resulted in larger error. Huelser and Metcalfe (2012) similarly did not observe a pretesting effect for the unrelated condition. These results did not support the error correction hypothesis because the condition with larger semantic error produced a null effect of pretesting, but the condition with smaller semantic error produced a significant effect of pretesting.

While these experiments had design issues (see Potts and Shanks, 2014), Seabrooke et al. (2022) designed experiments specifically to address these issues and largely arrived at the same conclusion. Seabrooke et al. (2022) designed a paradigm specifically to investigate the error correction hypothesis, which conceptually replicated the aforementioned results contradicting the error correction hypothesis. Seabrooke et al. were interested in probing the prediction that the size of errors in pretesting dictates the amount of learning, the key prediction of the error correction hypothesis, but they wanted to control for learning difficulty between low and high error conditions. To address this, Seabrooke had participants learn Finnish-English word pairs coming from two categories (four-legged animals and clothing). During test trials, participants guessed the category of the Finnish word, then guessed its English translation, followed by feedback with the correct target. In the read condition, participants studied the word pair for the whole trial duration. When participants correctly guessed the category, error was presumed to be lower than when they guessed the incorrect category. Both pretesting conditions led to greater retention than simply reading the word pairs, however, incorrect category guesses were not associated with greater retention than correct category guesses. Therefore, the critical assessment of the error correction hypothesis (whether error would increase retention) was not observed.

Although the extant evidence does not favor the error correction hypothesis (Grimaldi & Karpicke, 2012; Huelser & Metcalfe, 2012; Seabrooke et al., 2022), it is important to note that the error correction hypothesis differs from the EDL hypothesis, as conceptualized here. The error-correction hypothesis emphasizes the importance of overt errors. By contrast, EDL depends on the comparison between anticipated and actual targets, and learning can occur even when the correct target is produced. For instance, a neural network model of memory might generate an output that closely resembles a target pattern, but it could still learn to generate a more precise representation of the target on a subsequent trial. Within the EDL hypothesis, learning is maximized not by errors per se, but by stress testing the learner's internal representations of the study material. That is, if a learner were to correctly recall a target item after initially struggling, EDL should still take place, enabling them to more efficiently generate the targeted information on subsequent trials.

The EDL hypothesis is consistent with an arguably overlooked computational model developed by Mozer et al. (2004) as a follow up to Carrier and Pashler (1992). Mozer et al. developed a "complete cue processing" model based on error-driven learning. Their model attempted to predict targets from cues based on an error-driven learning mechanism known as the delta rule (Widrow & Hoff, 1960). Their implementation assumes that learning is not instantaneous and that cue processing, and therefore the formation of a prediction, ceases as soon as the target is processed. Accordingly, conditions that present the cue before the target allow a complete prediction to be formed and therefore optimize learning through error. They trained their model using Carrier and Pashler's (1992) results, producing quantitative predictions for retention as a function of SOA. The model predicted that cue presentation of 1 second would result in a 3% retention benefit, whereas cue presentation of 5 seconds would result in a 7%

benefit of testing. To our knowledge, there have been no behavioral replications of Carrier and Pashler's experiment with cue presentation times of fewer than five seconds.

Consistent with Mozer et al.'s complete cue processing model and the EDL hypothesis, Kornell, Klein, and Rawson (2015) argued that retrieval success or failure does not matter as long as a retrieval attempt is made and feedback is provided. Kornell et al. had participants study paired associates, then engage in cued recall. Critically, incorrect items were of interest and only incorrect items were randomly assigned to two experimental conditions. This procedure ensured similar difficulty between the two conditions. In the copy condition, participants copied the correct target. In the retrieval condition, participants retrieved the target successfully with the aid of the correct target word stem. On a criterial cued recall test, there were no differences observed between the copy and retrieval conditions, indicating that as long as an initial retrieval attempt was made, the source of correct feedback was irrelevant; it could be retrieved or exogenously presented. Kornell et al. showed that the retrieval attempt, not retrieval success or failure, is critical to retrieval-based learning. Moreover, Kornell et al. emphasized the importance of controlling for learning difficulty in investigations of retrieval-based learning mechanisms. In the pretesting experiments described earlier, where the size of errors was operationalized as the semantic distance between guesses and correct targets, the interpretation through the lens of EDL is that more learning almost certainly did take place; however, there was also more to learn for these conditions, giving rise to the lower performance for high error conditions compared to low error conditions.

Based on our goal to directly test the EDL hypothesis, we sought methods to evoke error during the retrieval attempt, without producing overt errors of commission (guesses). We developed a paradigm that bears significant resemblance to Carrier and Pashler (1992) and to

Vaughn, Hausman, and Kornell (2017). In the latter study, the authors predicted that increasing the retrieval attempt duration would increase retention. They manipulated the stimulus onset asynchrony (SOA) between the cue and the correct answer feedback, similarly to Carrier and Pashler (1992). In the "test" conditions, participants attempted to retrieve and guess the answers to general knowledge questions for 5, 10, or 30 seconds until feedback was provided. Vaughn et al. (2017) replicated the pretesting effect in that retention was greater for "test" SOAs (5, 10, and 30 seconds) than in the "study" SOA (0 seconds) condition. However, there were no significant differences in retention among "test" conditions, contrary to their predictions.

The manipulation of SOA presents an exciting opportunity to probe the EDL hypothesis, but in order to disambiguate its predictions from those of the error correction hypothesis, guessing and associated error correction must be eliminated. Furthermore, if such an association between SOA and retention exists, neurophysiology studies suggest that the relevant timescale may be much shorter than 5 seconds or more. For example, studies that used EEG recording during incidental word encoding contrasted neural activity associated with subsequently remembered words and subsequently forgotten words, revealing an event-related potential effect known as the late positive complex from 400 ms to 800 ms (e.g. Paller, Kutas, and Mayes, 1987). This result suggests that critical memory formation processes take place during the first 1,000 ms after stimulus presentation. More recently, Zhang, Fell, and Axmacher (2018) showed that a rapid sequence of neural firing known as a "ripple" that was recorded in human epilepsy patients 500 to 1,200 ms after picture presentation was "replayed" during non-REM sleep and that replay predicted subsequent memory. Taken together, these studies suggest that critical memory formation processes occur during the first 1 to 2 seconds of encoding.

Here, we report the results from four experiments designed to test the EDL hypothesis. We manipulated the amount of learning error in each trial by varying the cue-target stimulus-onset asynchrony (SOA). We probed the relationship between SOA and retention during foreign language vocabulary learning on the order of 200 milliseconds to 6 seconds. There are two distinct possibilities with respect to the relationship between SOA and retention in the current experiments. Consistent with the EDL hypothesis, "gap" SOAs–conditions where SOA is greater than 0 ms–are expected to result in greater error and therefore greater retention than the "no gap," 0 ms SOA condition. In contrast, the error correction account predicts that, in the absence of overt errors, there will not be differences between the gap and no gap conditions. Neither account makes specific claims about the specific gap SOAs that might produce greater retention than the no gap condition; however, the literature reviewed above suggests that an effect of SOA, if any, will be observed between 400 ms and 5 seconds. Moreover, Mozer et al.'s (2004) computational model predicted that any SOA greater than 0 seconds would be associated with greater retention than studying.

There were two phases: a learning phase in which participants who were naive to Swahili learned Swahili-English vocabulary pairs and a cued-recall final test phase. In all four experiments, total trial duration was fixed, such that an increase in SOA resulted in a complementary decrease in feedback duration. Critically, participants were instructed to attempt to retrieve the target word, but if they could not, they could wait until the target was presented. With these instructions, we intended that overt retrieval errors would be precluded almost entirely. In Experiments 1 and 2, we explored various SOAs in search of the error thresholds (if they exist) between learning through study and learning through retrieval. In Experiment 3, we used a blocked design to test whether increased attention or distinctiveness modulated the effect

of SOA. In Experiment 4, we manipulated the retention interval at half a minute and one day. Across all experiments, we predicted that performance on the final cued-recall criterial test would be enhanced in the gap relative to the no gap condition.

Experiment 1 Introduction

Experiment 1 was designed to probe the parameter space of SOA and retention with the intention of evaluating the EDL hypothesis. We expected that SOAs greater than 0 ms, the gap conditions, would cause greater retention than the no gap condition. We used shorter SOAs than Vaughn, Hausman, and Kornell (2017) due to their null finding of SOA using SOAs from 5 seconds to 30 seconds and neurophysiological evidence suggesting that encoding and retrieval processes are most pronounced during the first two seconds of processing.

Experiment 1 Methods

Experiment 1 Participants

Fifty-eight participants from the University of California, Davis online paid subject pool participated in part one and forty-five participants participated in part two in exchange for five dollars for each part. Seven participants were excluded from all analyses due to failing attention checks or reporting prior familiarity with the Swahili language, leaving 38 participants in the final sample. Given the novelty of our paradigm, we could not reliably anticipate the effect size. Rowland's (2014) meta-analysis of the testing effect literature found a mean effect size for paired-associates of 0.69. In the absence of an initial study phase, we reasoned that a smaller effect size of 0.5 may be expected. A power analysis in GPower (Faul et al., 2009) for a difference in means between matched pairs with power set at 0.80 and alpha = 0.05 indicated at least 34 participants were required to detect an effect size of Cohen's d = 0.5.

Experiment 1 Materials

Twenty Swahili-English word pairs were chosen from Nelson and Dunlosky's (1994) normative dataset. The Swahili-English pairs were sorted by highest recall performance after three study-test cycles. After excluding "rafiki-friend", the top twenty pairs were selected as experimental pairs and the next six were selected as practice pairs.

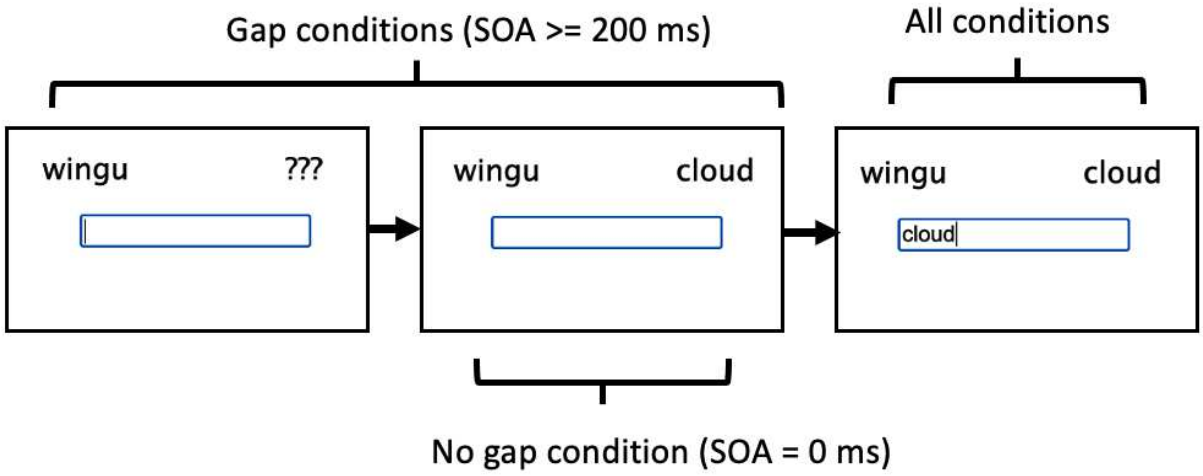

Figure 1.1 Experimental Design. In "gap" conditions, participants viewed a cue word until the SOA elapses at which point the target was presented. In "no gap" conditions, the cue and the target were presented for the whole trial. In all conditions, participants could type in the target at any point during the trial, but in gap conditions they were not expected to do so.

Experiment 1 Design

In this study, participants learned Swahili-English word pairs and completed a cued-recall test twenty-four hours later. The manipulation of error was implemented by varying SOA within participants. For each participant, the 20 experimental pairs were randomly assigned to one of the five following SOAs: 0, 1,500, 3,000, 4,500, and 6,000 milliseconds. SOA of 0 was considered a "no gap" trial and all other SOAs were considered "gap" trials.

Experiment 1 Procedure

After completing the informed consent, participants read the instructions and completed forced-choice instructions checks. Participants were instructed to type in the English translation for each Swahili word as soon as possible, and, as they learned the word pairs, they could start typing even before the English word was presented. Critically, there was no difference in behavioral task demands between study and test trials. The 20 word pairs were randomly assigned to the SOAs such that each SOA had four pairs. After completing six practice trials, participants began learning the experimental pairs which were presented in a random sequence. For all conditions, each trial was 8 seconds and there was a 1-second inter-trial interval filled with a fixation cross at the center of the screen. For trials of SOA of 0 ms, both the cue and target were presented simultaneously for the entire trial (cue-target). For trials with an SOA greater than 0 ms, the cue, but not the target (cue-???), was presented for the SOA duration. After the SOA elapsed, the target was presented in addition to the cue (cue-target). Both the cue and target remained on the screen until the remaining trial time elapsed. For all conditions, a text box was available to type in the English translation during the entire 8 seconds or until a response was typed and the 'enter' key was pressed. After each of the twenty pairs had been presented, the pairs were presented again in a new random sequence, and this process was repeated until five total repetitions had been completed. Twenty-four hours after the first phase began, participants completed a cued-recall final test. The Swahili cue words were presented one at a time and the test was self-paced. There was no back button and the sequence of the cues was randomized. Finally, participants indicated whether they had any experience with Swahili prior to the current experiment.

Experiment 1 Analysis Methods

Learning Accuracy and RT Analysis

In order to validate our manipulation of SOA, we first examined reaction time (RT) during the learning phase. Due to the long and variable durations of word typing, the first key press was used to record RT. In the RT Model, we examined RT during the learning phase using a two-way, within-subjects ANOVA to examine effects on RT by stimulus-onset asynchrony (SOA) and learning repetition. We used R (version 3.6.3, R Core Team, 2020) and the afex library function aov_ez to conduct our ANOVA using type three sum of squares (Singmann et al., 2022).

Next, we explored target accuracy in order to probe whether participants guessed targets (i.e., most responses were incorrect) as in pretesting paradigms and whether participants achieved a similar amount of learning in each SOA condition. Target accuracy scoring was strict, with only a perfect match to the correct target being scored as correct. In the Learning Accuracy Model, we used a two-way, within-subjects ANOVA to examine effects on accuracy by stimulus-onset asynchrony (SOA) and learning repetition.

Cued-Recall Analysis

We employed two models to characterize the effect of SOA on retention and test our hypotheses. In the No gap/Gap Model, we coded the 0 ms SOA condition as "no gap" and the four SOAs of 1,500, 3,000, 4,500, and 6,000 ms were coded as "gap". We created a generalized linear mixed-effects model (Jaeger, 2008) with a binomial distribution and a logistic link function, cued-recall accuracy as the outcome variable, a fixed effect of learning condition (no gap or gap), a random slope for the effect of learning condition within participants, and random intercepts for participant and word pair (accuracy ~ NoGap_Gap + (NoGap_Gap | id) + (1 | pair_idx)). We included a random intercept for word pairs to account for variance in the learnability of individual word pairs and we included the random slope to account for differently

sized effects of learning condition (Barr et al., 2013). The generalized linear mixed-effects model was fit using the afex library's function 'mixed' (Singmann et al., 2022), which used lme4's 'glmer' function (Bates et al., 2015).

We used a parametric bootstrap, type III sum of squares method for predictor inference which was recommended for experiments similar to the present ones (Singmann & Kellen, 2019). There are difficulties in inference for generalized mixed effects models due to the inability to estimate denominator degrees of freedom, and inflated type 1 errors are associated with likelihood ratio tests, particularly when random factors have fewer than 40 levels (Pinheiro & Bates, 2000). Therefore, predictor inference was carried out with the pbkrtest library's 'PBmodcomp' function. In a similar way to a likelihood-ratio test procedure, PBmodcomp compares a reduced model to the full model in order to make inferences about a left out predictor. The PBmodcomp parametric bootstrap procedure simulates datasets from the reduced model, then fits both the reduced and the full model to each dataset. The parametric bootstrap p-value corresponds to the percentage of simulated likelihood-ratio values that are larger than the observed likelihood-ratio value (Halekoh & Højsgaard, 2014). We used R (version 4.2.2, R Core Team, 2022) to compute all analyses.

In the Categorical Model, we used distinct predictors for each SOA (0, 1,500, 3,000, 4,500, and 6,000 ms). A random slope was not included in this model due to convergence warnings (accuracy ~ SOA + (1 | sonida_id) + (1 | pair_idx)). All statistical computations were identical to the first model, except that instead of leaving out the predictor SOA, each individual parameter was left out, allowing inference for each level of SOA. Given the issues described above, we believe this to be the most robust method to calculate p-values and for these data.

There was no preregistration for this experiment or subsequent experiments. All study materials, data, and code for this experiment and all subsequent experiments can be found at https://github.com/wbreilly/error_driven_learning_vocabulary. This study received ethics committee approval from the UC Davis Institutional Review Board.

Experiment 1 Results

SOA	mean	sd
0	1,376	413
1500	2,208	617
3000	3,013	785
4500	3,659	872
6000	4,579	1,084

SOA	mean	sd
0	0.77	0.38
1500	0.77	0.38
3000	0.74	0.36
4500	0.67	0.37
6000	0.48	0.34

Table 1. Mean and standard deviation of reaction time for each condition in Experiment 1.
Table 2. Mean and standard deviation of target accuracy for each condition in Experiment 1.

Learning accuracy was higher and reaction times were faster for shorter SOA conditions

In order to validate our manipulation of SOA, we first examined reaction time (RT) during the learning phase. For longer SOAs, we would expect RTs to be slower because participants have to wait longer to learn the correct target. Note that due to the long and variable durations of word typing, the first key press was used to record RT. We examined RT during the learning phase using a two-way, within-subjects ANOVA to examine effects on RT by stimulus-onset asynchrony (SOA) and learning repetition. The RT ANOVA revealed significant effects of SOA, learning repetition, and an interaction (all Fs > 20, all ps < .001). By the end of the

learning phase, RT was less than the SOA for the SOAs of 3000, 4500, and 6000, which is suggestive of successful retrieval on average. Furthermore, RT accelerated across repetitions, RT was slower for longer SOAs, and that RT acceleration was greater for longer SOAs. These results provide solid evidence SOA manipulation impacted the duration of retrieval attempts in that longer SOAs were associated with longer RTs.

Next we explored target accuracy in order to ensure that participants were waiting for the correct target to be presented to them and to verify that participants achieved a similar amount of learning in each SOA condition. We used a two-way, within-subjects ANOVA to examine effects on accuracy by stimulus-onset asynchrony (SOA) and learning repetition. The Learning Accuracy model revealed a significant effect on target accuracy of SOA ($F = 37.89$, $p < .001$) and a significant interaction of SOA and learning repetition ($F = 7.14$, $p < .001$). We concluded that the early decreased accuracy in the 6000 SOA condition was likely due to having less time to type the full target word correctly (2 seconds compared to as much as 8 seconds in the 0 SOA "study" condition), therefore in future experiments, we provided more time in the feedback part of each trial. The RT and accuracy data together indicate that participants were capable of learning the word pairs to a similar degree and that the manipulation of SOA resulted in the expected differences in the duration of retrieval attempts.

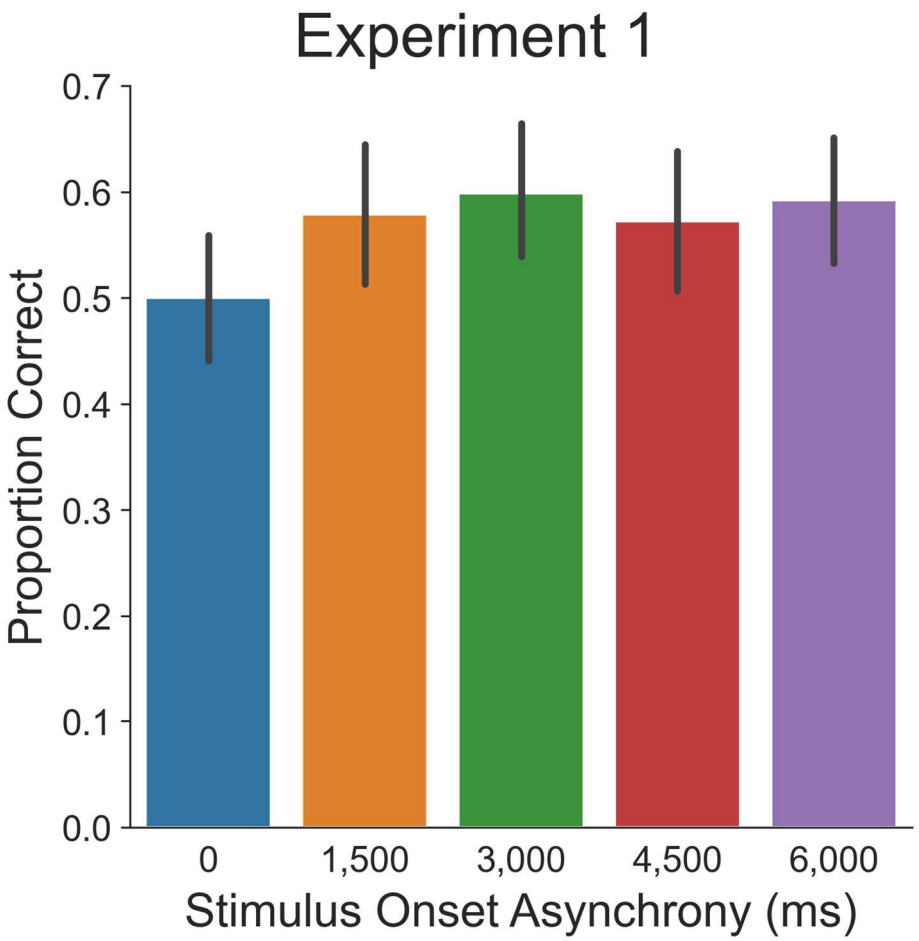

Figure 1.2. Experiment 1 Cued-recall Performance. Mean proportion correct on the final cued-recall test in Experiment 1. Error bars represent standard error of the mean.

Retention effect of gap compared to no gap in cued-recall

In the critical analysis of final test performance, we predicted that the gap SOAs would show greater retention than the no gap condition. To begin our investigation of the impact of SOA on cued recall, we coded the 0 ms SOA condition as "no gap" and the four SOAs of 1,500, 3,000, 4,500, and 6,000 ms were coded as "gap". As described in the methods section, we created a generalized linear mixed-effects model with cued-recall accuracy as the outcome variable, a fixed effect for learning condition (gap or no gap), a random slope for learning condition within participant, and random intercepts for participant and word pair. The parametric

bootstrap procedure revealed a significant effect of learning condition (X^2 = 4.16, p = .045), indicating that learning with "gap" (M = 0.59, SD = 0.37) resulted in greater retention than learning with "no gap" (M = 0.50, SD = 0.37). To break down this effect, we created a second model that tested whether exclusion of each gap level of SOA (1,500, 3,000, 4,500, and 6,000 ms) decreased model fit utilizing the same parametric bootstrap procedure, revealing significant effects on retention for 3000 ms (X^2 = 4.48, p = .041) and 6000 ms (X^2 = 4.62, p = .033).

Experiment 1 Discussion

Results from Experiment 1 were consistent with our prediction that simply including a gap between cue and target would be sufficient to increase retention relative to a pure restudy condition. Moreover, participants had very high target accuracy, indicating that guesses and error correction were largely absent, as in pretesting paradigms. This effect supported the EDL hypothesis in that the SOAs that provided time for a retrieval attempt were associated with greater retention than the SOA of 0 ms. We expected a gap effect for each condition but this was supported by descriptive statistics only in the 1,500 and 4,500 gap conditions. One possible explanation is that individual SOAs had insufficient power in only having four word pairs each. Subsequent experiments assigned more word pairs to each SOA. Taken together, these results suggest a cue-target gap was sufficient to enhance final test performance. In Experiment 2, we investigated whether even shorter SOAs might be sufficient to enhance learning.

Experiment 2 Introduction

Experiment 2 used essentially the same design as Experiment 1, with the principal difference being that Experiment 2 used shorter SOAs.

Experiment 2 Methods

Forty-five participants from the University of California, Davis online paid subject pool participated in part one and 37 participants participated in part two in exchange for $5 for each part. Eight participants were excluded from all analyses due to failing attention checks or reporting prior familiarity with the Swahili language, leaving 29 participants in the final sample.

Experiment 2 methods were fundamentally the same as in Experiment 1 except for the following changes. The total trial time was shortened to 4000 ms and the SOAs were 0, 200, 600, and 800. In order to maintain a similar study phase duration to Experiment 1 and to increase power, we used 28 Swahili-English word pairs selected in the same way as in Experiment 1. In Experiment 2, there were four repetitions of each word pair during the learning phase.

Experiment 2 Results

SOA	mean	sd
0	1,115	253
200	1,182	211
600	1,421	151
800	1,571	176

SOA	mean	sd
0	0.84	0.26
200	0.84	0.27
600	0.81	0.25
800	0.82	0.26

Table 3. Mean and standard deviation of reaction time for each condition in Experiment 2.
Table 4. Mean and standard deviation of target accuracy for each condition in Experiment 2.

No significant differences in learning accuracy, mean RTs were longer than SOAs

The learning phase of Experiment 2 was analyzed in an identical manner to Experiment 1. The RT model revealed significant main effects of SOA and repetition and the interaction was not significant. There were smaller differences between SOAs than in Experiment 1 reflecting the reduction in SOA variance in Experiment 2. The mean RT at the end of learning was greater than 1 second for all SOAs, indicating that participants did not respond before target presentation on average, unlike in a retrieval practice paradigm. Nevertheless, the difference between mean RT in the longest SOA condition, 800 ms, and the shortest, 0 ms, was about 500 ms, suggesting that participants responded 300 ms faster than if they had passively waited for the target to be presented in the 800 ms condition.

The Learning Accuracy model revealed no significant differences in SOA or in repetition. Mean target accuracy was around .80 or greater for all conditions indicating that participants had ample time to type in the correct target after target presentation and that they were not making guesses.

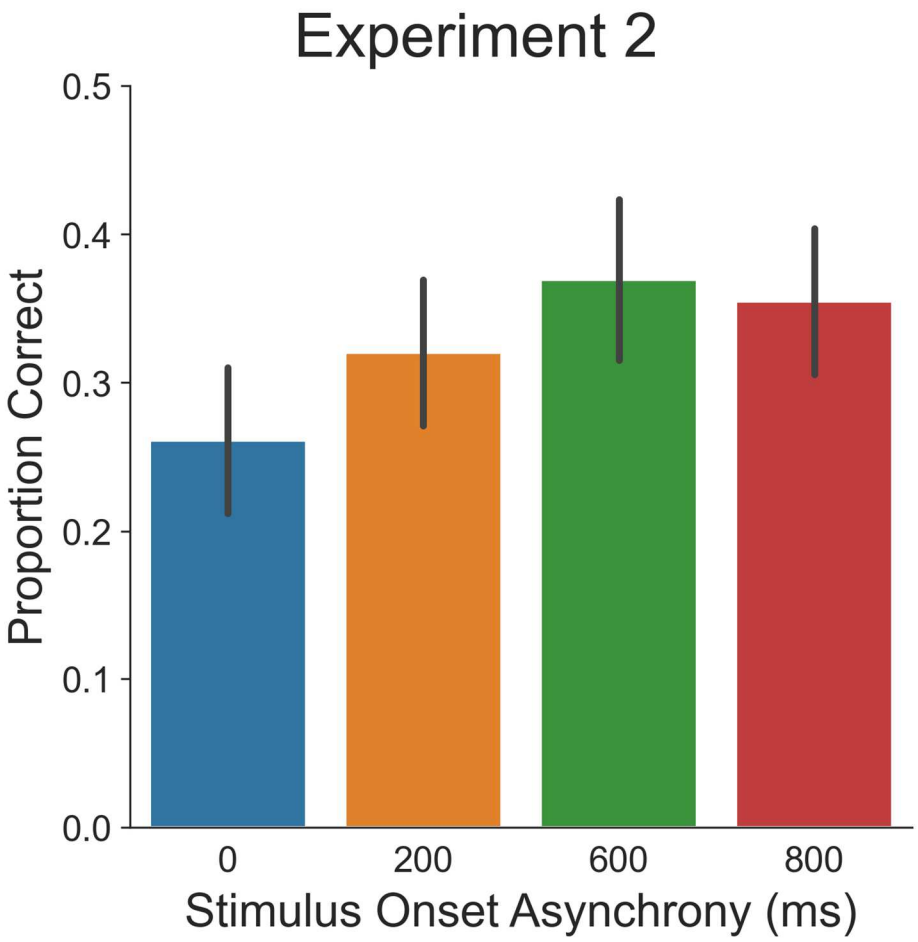

Figure 1.3. Experiment 2 Cued-Recall Performance. Mean proportion correct on the final cued recall test in Experiment 2. Error bars represent standard error of the mean.

600 and 800 ms gaps resulted in greater retention than no gap

Our analysis of the cued recall final test was fundamentally the same as in Experiment 1. In the No gap/Gap model, the 0 ms SOA condition was coded as "no gap", and the three SOAs of 200, 600, and 800 ms were coded as "gap". This model revealed that learning with a "gap" (M = .35, SD = .25) resulted in significantly greater retention (X^2 = 7.52, p = .004), than learning with "no gap" (M = .26, SD = .26). This result conceptually replicated the effect of gap compared to no gap observed in Experiment 1, but with a four second trial duration and much shorter SOAs. The Categorical model revealed significant effects on retention for 600 ms (X^2 =

9.39, p = .008) and 800 ms (X^2 = 5.51, p = .032), but the 200 ms SOA was not significant (X^2 = 2.06, p = .151). This remarkable result indicates that delaying target presentation by as little as 600 ms is sufficient to provide a significant retention benefit 24 hours later compared to simply studying.

Experiment 2 Discussion

Experiment 2 replicated the retention benefit of gap over no gap with much shorter SOAs. There was a significant difference in retention one day after inserting a 600 ms gap between cue and target presentation over four repetitions. These results are consistent with the neural data indicating that critical memory formation processes take place in the first second after stimulus presentation. The next experiments included SOAs of 200, 400, and 600 ms. A significant effect of gap at 400 ms would dovetail with the occurrence of the LPC at 400 ms.

Experiment 3 Introduction

Experiments 1 and 2 established the basic phenomenon that gap SOAs were associated with greater retention than no gap. To investigate this phenomenon further, Experiment 3 was designed to test whether the mixed list design of the first two experiments may have contributed to the effects we observed. One reason could be that attention or other cognitive resources were diverted from no gap trials to gap trials. Alternatively, a well-known mnemonic principle is that manipulations that enhance stimulus distinctiveness enhance retention in mixed lists of control and distinct items, but not when the same items are studied in separate lists (e.g., Waddill & McDaniel, 1998). Accordingly, in Experiment 3 learning was separated into two separate blocks each containing word pairs of the same condition, either gap or no gap. If the results reveal a

retention benefit for gap compared to no gap, even in this design, then it would appear unlikely that the effect of gap is an artifact of mixed-list learning.

Experiment 3 Methods

Instructions and the procedure of individual trials in Experiment 3 were identical to those of the first two experiments. The only difference was that the SOA manipulation was no longer fully within-subjects and no longer in a mixed-list design. Learning condition was manipulated within-subjects such that each participant had one block of no gap and one block of gap, therefore there were two total blocks of learning trials. The three gap SOAs (200, 400, or 600 ms) were manipulated between subjects such that each participant was randomly assigned one block of study, and one block of one of the three possible test SOAs, and block order was randomly assigned. This experiment saw disproportionate dropout further complicated by online data collection therefore the final number of participants was not equalized. In the "gap first" order, there were 17, 18, and 29 participants in the 200, 400, and 600 ms gap conditions, respectively. In the "gap second" order, there were 16, 18, and 29 participants in the 200, 400, and 600 ms gap conditions, respectively.

Experiment 3 Results

SOA	mean	sd
0	1,078	299
200	1,137	219
400	1,212	144
600	1,413	220

SOA	mean	sd
0	0.89	0.23
200	0.90	0.21
400	0.93	0.17
600	0.84	0.28

Table 5. Mean and standard deviation of reaction time for each condition in Experiment 3.
Table 6. Mean and standard deviation of target accuracy for each condition in Experiment 3.

No significant differences in learning accuracy between groups, RTs were longer than SOA

The Experiment 3 learning phase was analyzed with the same purpose as the two previous experiments, to ensure expected results of similar accuracy between conditions and that SOA affected RT. In Experiment 3, distinct groups of participants (gap group) were assigned only one "gap" SOA (200, 400, or 600), but all groups had the no gap 0 ms SOA. To accommodate this design, SOA was coded as gap or no gap as a within-participants factor, and gap group was a within-participants factor. As expected, the RT model revealed significant differences in gap group, gap/no gap, a significant interaction of the two, and a significant effect of repetition.

The learning accuracy model included the same variables as the RT model, revealing only a significant difference in repetition. Because different participants completed different gap conditions, these results showed that any differences in retention as a function of SOA were not due to differences in learning accuracy.

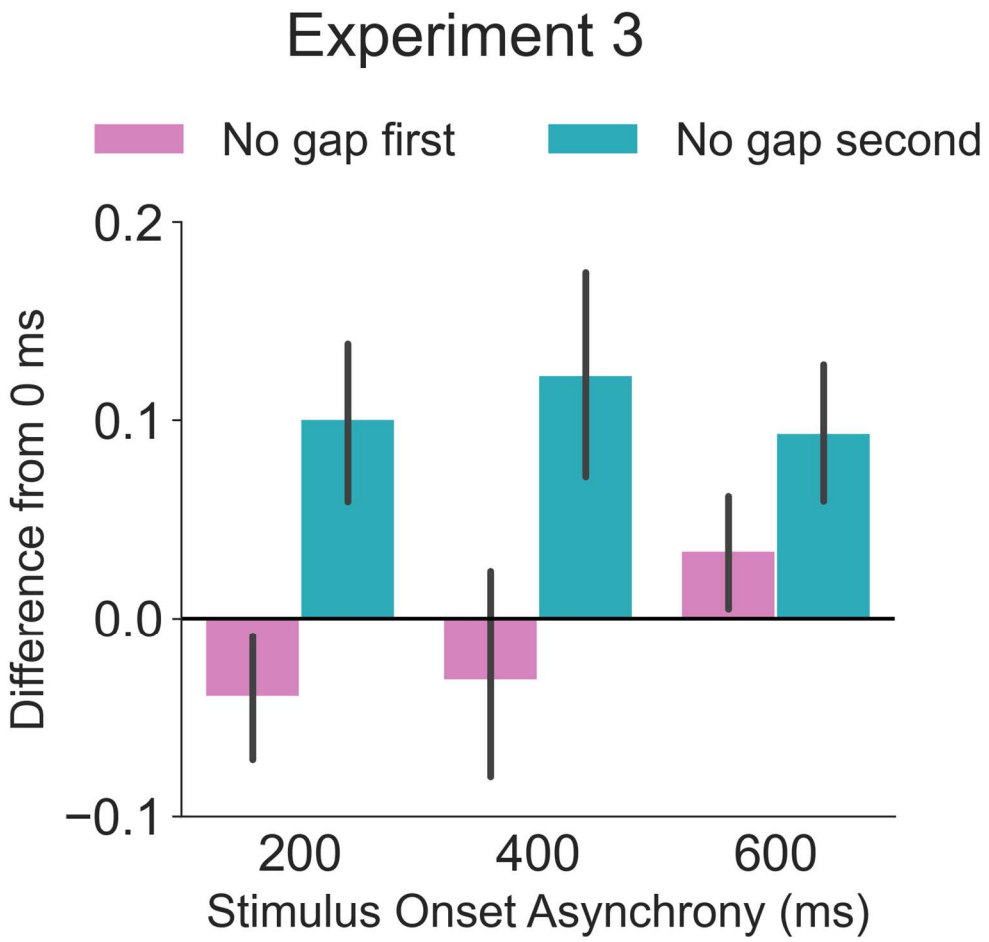

Figure 1.4. Experiment 3 Cued-Recall Performance. Mean difference in proportion correct in the gap condition compared to no gap and separated by which condition was learned first on the final cued recall test in Experiment 3. In Experiment 3, the gap conditions and order of gap and no gap were between participants, but each group had the no gap condition. Error bars represent standard error of the mean.

The 600 ms condition was robust to a pure list design and primacy effects

Once again, we took a very similar analysis approach for the final test in Experiment 3. To accommodate the design differences and anticipated primacy effects, a between-participants factor for block order was added to each model. In the No Gap/ Gap model, the 0 ms SOA condition was coded as "no gap", and the three SOAs of 200, 600, and 800 ms were coded as "gap". This model revealed a significant fixed effect of gap compared to no gap (X^2 = 6.84, p = .016), and a significant interaction with block order (X^2 = 11.79, p = .002). In order to break

down this effect, the categorical model compared the three gap SOAs to the no gap SOA. This model revealed significant effects of 200 ms (X^2 = 6.79, p = .016), 400 ms (X^2 = 15.61, p = .003), and 600 ms (X^2 = 13.49, p = .003), and significant interactions between block order and 200 ms (X^2 = 6.13, p = .027) and 400 ms (X^2 = 11.53, p = .003). The interaction between 600 ms and block order was not significantly different (X^2 = 2.95, p = .093). These results and the mean differences plotted in Figure XX demonstrate that the gap conditions were more resilient to primacy effects than no gap overall, but only the 600 ms condition showed greater retention than the no gap condition in both block orders.

Experiment 3 Discussion

Experiment 3 tested whether retention benefits of gap compared to no gap would replicate in a design that separated the two conditions into separate blocks of learning trials. The results suggested a primacy effect in that the benefit of gap compared to no gap was greatest when the gap conditions were in the first block of learning. Despite this, the 600 ms condition revealed numerically greater retention than the no gap condition regardless of block order. These results provide solid evidence that the effects on retention of inserting a gap between cue and target presentation in Experiments 1 and 2 are not due to the use of mixed list designs.

Experiment 4 Introduction

In Experiments 4a and 4b, we sought to replicate our findings and probe whether retention interval is a boundary condition. Some studies have shown that the retention benefits of testing emerge only after a long retention interval (e.g., Roediger and Karpicke, 2006); however, Rowland's (2014) meta-analysis revealed reliable benefits of testing even at short retention intervals. Moreover, the effect size of testing compared to study was shown to increase with

retention interval (Rowland, 2014). The design of these experiments was very similar to Experiment 2, with the modifications the gap SOAs were the same 0, 200, 400, and 600 ms and Experiment 4b used a 30 second retention interval. The EDL hypothesis does not make specific predictions about retention interval, therefore we expected to observe an effect of gap compared to no gap in both the short and long delay groups.

Experiment 4 Methods

Seventy-one participants recruited from Amazon Mechanical Turk completed parts one and two in the one-day delayed group in exchange for $4. One hundred participants recruited from Amazon Mechanical Turk completed parts one and two in the 30-second delayed group in exchange for $3. Seventy-eight participants were excluded from all analyses for failing data quality checks, leaving 44 participants in the one-day delay group and 49 participants in the one-minute delay group.

Experiment 4 methods were fundamentally the same as in Experiment 1 except for the following changes. The total trial time was still 4000 ms but the SOAs were 0, 200, 400, and 600 ms, as in Experiment 3. SOA was manipulated within participants. Two separate groups of participants completed an identical learning phase, then completed an identical final cued recall test. In Experiment 4a, participants completed the final test 24 hours after learning, as in the previous experiments. In Experiment 4b, participants completed the final test 30 seconds after learning.

Experiment 4 Results

Learning accuracy and RT control analyses replicated previous experiments

The learning phase of Experiment 4a was analyzed in an identical manner to the previous experiments. The RT model revealed significant main effects of SOA and repetition and the interaction was not significant. The Learning Accuracy model revealed no significant differences in SOA or in repetition.

The learning phase results for Experiment 4b were very similar to Experiment 4a. The RT model revealed significant main effects of SOA and repetition and the interaction was not significant. The Learning Accuracy model revealed no significant differences in SOA or in repetition.

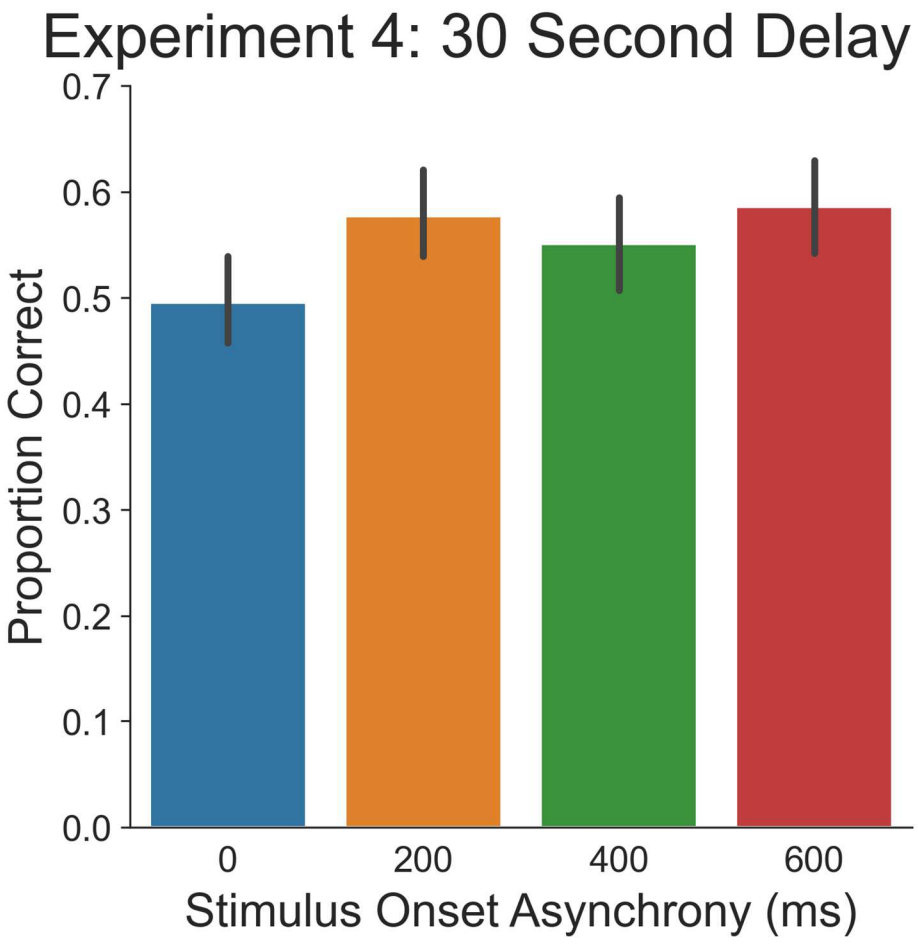

Figure 1.5. Experiment 4 30s Delay Cued-Recall Performance. Mean proportion correct on the final cued recall test in Experiment 4, 30 second delay group. Error bars represent standard error of the mean.

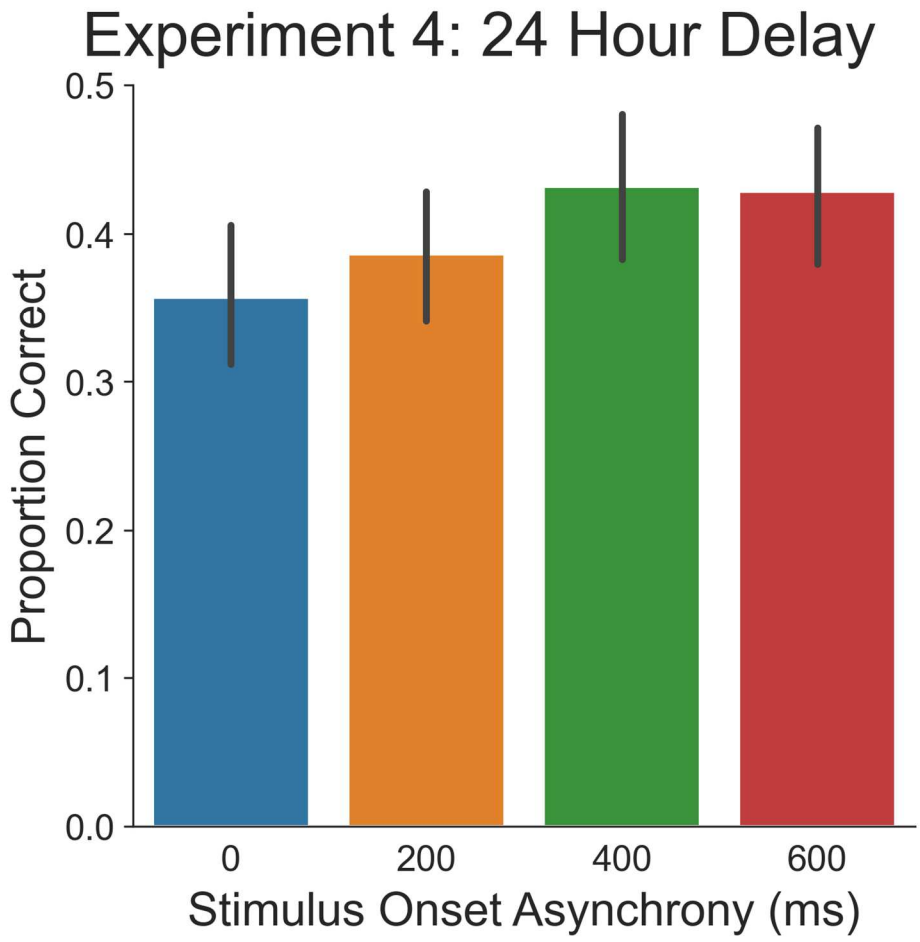

Figure 1.6. Experiment 4 24-hour Delay Cued-Recall Performance. Mean proportion correct on the final cued recall test in Experiment 4, 24 hour delay group. Error bars represent standard error of the mean.

The 600 ms condition produced a retention benefit after a 30-second and one-day retention interval

In the cued recall analysis, we once again took a similar approach, with the exception that a fixed effect of retention interval was added to all models. The No gap/Gap Model revealed that retention of gap pairs was significantly different from no gap pairs (X^2 = 7.92, p = .008) and a

fixed effect of retention interval revealed that retention was greater at the 30 second delayed test(X^2 = 7.57, p = .011), but the interaction was not significant (X^2 = 0.38, p = .50). In the categorical model, there was a significant effect of 600 ms (z = 4.38, p = .032) and a significant effect of retention interval (X^2 = 5.58, p = .025). All other effects and interactions were not significantly different from zero (all X^2 < 2.6, all p's > .14).

Experiment 4 Discussion

Experiment 4 was designed to test whether the gap effect was contingent on a long retention interval (24 hours), or if it would also be observed at a short retention interval (30 seconds). The results once again revealed that inserting a gap between the cue and target resulted in a retention benefit. The 600 ms gap revealed a significant benefit over No Gap at an immediate test, and replicated the effect of 600 ms in Experiment 2 and 3. These results also replicated Carrier and Pashler's (1992) Experiment 2 in that they observed a gap effect at a short and long retention interval, and there was no interaction between gap condition and retention interval.

Chapter 1 General Discussion

The results of four experiments reported here support the error-driven learning hypothesis in that—in the absence of overt errors—a gap between cue and target resulted in greater retention than studying. In Experiment 1, we showed that inserting a gap between cue and target presentation during learning of foreign language vocabulary enhanced retention one day later compared to the no gap condition. Experiment 2 provided evidence that a gap of as short as 600 ms was sufficient to produce a retention effect compared to the no gap condition. Experiment 3 ruled out the alternative explanation that the effect of gap was an artifact of a mixed-list design.

Experiment 4 showed that the effect of a 600 ms gap was robust at both 30-second and one-day retention intervals. In Experiments 2, 3, and 4, the SOAs were all shorter than 1 second, reaction times during learning were longer than SOAs, and accuracy was high, providing strong evidence that participants were not learning from overt errors, as required by the error correction account. Taken together, these results provide evidence consistent with the error-driven learning hypothesis.

The primary purpose of this study was to begin to investigate whether error-driven learning might account for retrieval-based learning effects. In order to test the EDL hypothesis, and to disambiguate it from the error correction hypothesis, it was necessary to develop a paradigm that encouraged participants to begin and end a retrieval attempt without making overt errors. The retention effects reported here cannot be accounted for by the error correction account because participants made very few errors during learning and reaction times were longer than SOAs in all but Experiment 1. To our knowledge, this is the first study to report a retention benefit of sub-five-second retrieval attempts, let alone sub-one-second retrieval attempts, compared to studying.

Our experimental paradigm was designed to push the notion that retrieval attempts are fundamental to retrieval-based learning—successful overt retrieval and errors were virtually impossible to complete before target presentation. Each gap trial was nearly identical to each no gap trial, except for the brief gap between cue and target presentation. In Experiments 2, 3, and 4, there was very little time for anything except starting the retrieval attempt before target presentation, and we assumed that the retrieval attempt ended at that time. In Experiments 2, 3, and 4, there were no significant differences in learning accuracy. Reaction times were slower as SOA increased, which is consistent with the idea that longer SOAs were associated with longer

retrieval attempts. In Experiments 2, 3, and 4, mean reaction time did not increase equivalently to SOA increases. This suggested that although participants were responding more slowly to gap trials than no gap trials, participants responded more quickly as a result of learning. Taken together, these results suggest that the cognitive processing afforded by a 600 ms gap is what led to the retention benefits observed here. Although the current study was focused on adjudicating between the error correction account and the EDL hypothesis, the EDL hypothesis is consistent with a retrieval effort hypothesis (e.g., Bjork, 1994; Pyc & Rawson, 2009) and the episodic context account (Karpicke et al., 2014). Future studies should aim to further distinguish between these accounts.

Using a similar design to the present one, Carrier and Pashler (1992) showed that a gap of 5 seconds resulted in superior retention to studying. Carrier and Pashler (1992) cited foundational error-driven learning papers to explain this effect, noting that retrieval practice involves error correction, whereas restudy "prevents the network from knowing what it would have produced on its own, and thereby inhibits it from properly correcting for any error." Their account emphasized the importance of correcting overt errors, whereas the more recent computational model from Mozer et al. (2004) emphasized the idea that incomplete episodic retrieval can be beneficial to learning. In their model, partial retrieval was operationalized by allowing that the memory system operates on "cycles" of computation in which a representation of a cue and a predicted outcome require several cycles to settle. Similarly, recent biologically-based computational models of the human hippocampus have shown that error-driven learning mediated by alternating circuits on a 200 ms cycle is able to outperform a "hebbian" hippocampus (Ketz et al., 2013) and are able to account for an array of retrieval-based learning phenomena (Liu et al., 2021; Liu & Ranganath, 2021; Zheng et al., 2022).

Our EDL hypothesis did not make specific predictions about the minimum SOA that would produce a retention effect or attempt to model the relationship between SOA and retention. The Mozer et al. (2004) complete cue processing model, however, simulated a 3% effect of a one second SOA, and an even smaller effect was predicted below one second. That model holds that increased cue processing time allows a more complete representation of the predicted target to form, therefore allowing more effective error-driven learning and greater retention. If that model is true, then it follows that after a certain amount of time, additional cue processing time is no longer effective. In contrast, one might expect that error accumulates continuously until the target is presented, therefore longer SOAs should be associated with larger retention benefits, but this was not supported in the current results or in Vaughn et al. (2017).

Our results also differed from a previous investigation using a pretesting paradigm with variable SOAs (Vaughn et al., 2017). Vaughn et al. concluded that the amount of time attempting retrieval had no effect on retention; however, their shortest SOA was five seconds. In contrast, we have shown that very short SOAs, less than 600 ms, have a null effect on retention, unlike the longer SOAs. Vaughn et al. (2017) controlled for SOA but manipulated feedback processing time at two or seven seconds and observed that seven seconds resulted in greater retention. Although we did not attempt to replicate this finding, results from Experiment 1 were not consistent with a tradeoff between retrieval attempt and feedback processing time. In all of our experiments, total trial duration was fixed, therefore, as SOA increased, feedback time decreased. If retrieval attempt duration does not matter, but feedback time does, then we would expect the greatest retention benefits for the shortest SOA conditions, and the least benefits for the longest SOA conditions. This pattern was not present in the current experiments. One possible explanation is that Vaughn et al. (2017) used trivia facts, which may have been more

conducive to elaboration, whereas our foreign language vocabulary words are less semantically rich. A further possibility is that the 600 ms SOA is long enough to form a prediction, but not so long that valuable feedback time is wasted. Future studies should develop the EDL hypothesis with empirical and simulated evidence to align these discrepant results.

One clear educational application of the current study is that students need not agonize over long retrieval attempts. Retrieval practice is recommended by cognitive scientists (Roediger & Karpicke, 2006; Roediger et al., 2011) supported by meta-analyses (Rowland, 2014) and meta-analyses of classroom studies (Hattie, 2008; Adesope et al., 2017), and widely employed by learning apps and educational products. Based on the current findings, teachers and learning product developers may provide feedback after as little as 600 ms, which is arguably more enjoyable for the learner and would allow more time for feedback or for additional trials. One limitation of the current study is that we did not directly manipulate feedback time or total trial time. Future studies should investigate these parameters to determine whether it is possible to achieve more learning in a given period of time. Moreover, unlike Carrier and Pashler's (1992) design and the testing effect paradigm, participants in current experiments did not experience an initial study phase. This is similar to the pretesting paradigm, except that we used multiple repetitions. Under the EDL hypothesis, an initial study phase before testing would not be expected to enhance retention compared to all testing. This question should be explored in future studies.

To conclude, we presented the EDL hypothesis in which we argued that retrieval-based learning capitalizes on an error-driven learning mechanism. We presented a novel paradigm and results that are consistent with the EDL hypothesis. To our knowledge, these results are the first to demonstrate a learning effect by attempting retrieval for less than one second.

Chapter 1 Works Cited

Adesope, O. O., Trevisan, D. A., & Sundararajan, N. (2017). Rethinking the Use of Tests: A Meta-Analysis of Practice Testing. Review of Educational Research, 87(3), 659–701. https://doi.org/10.3102/0034654316689306

Agarwal, P. K., Nunes, L. D., & Blunt, J. R. (2021). Retrieval Practice Consistently Benefits Student Learning: A Systematic Review of Applied Research in Schools and Classrooms. Educational Psychology Review, 33(4), 1409–1453. https://doi.org/10.1007/s10648-021-09595-9

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. Journal of Statistical Software, 67, 1–48. https://doi.org/10.18637/jss.v067.i01

Carrier, M., & Pashler, H. (1992). The influence of retrieval on retention. Memory & Cognition, 20(6), 633–642. https://doi.org/10.3758/BF03202713

Electrophysiological mechanisms of human memory consolidation | Nature Communications. (n.d.). Retrieved December 14, 2022, from https://www.nature.com/articles/s41467-018-06553-y

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

Full article: Retrieval attempts enhance learning regardless of time spent trying to retrieve. (n.d.). Retrieved December 14, 2022, from https://www.tandfonline.com/doi/full/10.1080/09658211.2016.1170152?casa_token=jO-HDj8dzoUAAAAA%3A_9uSuq9ITXVIsQq0xfApZwwUhP4bFec2Mt6ZS_C-ynRrZ4GVN5tsCrbLppkswDSF_SJgb-5XpRsy

Grimaldi, P. J., & Karpicke, J. D. (2012). When and why do retrieval attempts enhance subsequent encoding? Memory & Cognition, 40(4), 505–513. https://doi.org/10.3758/s13421-011-0174-0

Halekoh, U., & Højsgaard, S. (2014). A Kenward-Roger Approximation and Parametric Bootstrap Methods for Tests in Linear Mixed Models – The R Package pbkrtest. Journal of Statistical Software, 59, 1–32. https://doi.org/10.18637/jss.v059.i09

Hattie, J. (2008). : A Synthesis of Over 800 Meta-Analyses Relating to Achievement. Routledge. https://doi.org/10.4324/9780203887332

Hattie, J. (2011). : Maximizing Impact on Learning. Routledge. https://doi.org/10.4324/9780203181522

Hoppe, D. B., Hendriks, P., Ramscar, M., & van Rij, J. (2022). An exploration of error-driven learning in simple two-layer networks from a discriminative learning perspective. Behavior Research Methods, 54(5), 2221–2251. https://doi.org/10.3758/s13428-021-01711-5

Huelser, B. J., & Metcalfe, J. (2012). Making related errors facilitates learning, but learners do not know it. Memory & Cognition, 40(4), 514–527. https://doi.org/10.3758/s13421-011-0167-z

Jaeger, T. F. (2008). Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models. Journal of Memory and Language, 59(4), 434–446. https://doi.org/10.1016/j.jml.2007.11.007

Karpicke, J. D., Lehman, M., & Aue, W. R. (2014). Chapter Seven - Retrieval-Based Learning: An Episodic Context Account. In B. H. Ross (Ed.), Psychology of Learning and Motivation (Vol. 61, pp. 237–284). Academic Press. https://doi.org/10.1016/B978-0-12-800283-4.00007-1

Ketz, N., Morkonda, S. G., & O'Reilly, R. C. (2013). Theta Coordinated Error-Driven Learning in the Hippocampus. PLOS Computational Biology, 9(6), e1003067. https://doi.org/10.1371/journal.pcbi.1003067

Kornell, N., Hays, M. J., & Bjork, R. A. (2009). Unsuccessful retrieval attempts enhance subsequent learning. Journal of Experimental Psychology: Learning, Memory, and Cognition, 35, 989–998. https://doi.org/10.1037/a0015729

Kornell, N., Klein, P. J., & Rawson, K. A. (2015). Retrieval attempts enhance learning, but retrieval success (versus failure) does not matter. Journal of Experimental Psychology: Learning, Memory, and Cognition, 41, 283–294. https://doi.org/10.1037/a0037850

Liu, X. L., O'Reilly, R. C., & Ranganath, C. (2021). Chapter Four - Effects of retrieval practice on tested and untested information: Cortico-hippocampal interactions and error-driven learning. In K. D. Federmeier & L. Sahakyan (Eds.), Psychology of Learning and Motivation (Vol. 75, pp. 125–155). Academic Press. https://doi.org/10.1016/bs.plm.2021.07.003

Liu, X., Ranganath, C., & O'Reilly, R. C. (2022). A complementary learning systems model of how sleep moderates retrieval practice effects. OSF Preprints. https://doi.org/10.31219/osf.io/5aqwp

McClelland, J. L., & Rumelhart, D. E. (1981). An interactive activation model of context effects in letter perception: I. An account of basic findings. Psychological Review, 88, 375–407. https://doi.org/10.1037/0033-295X.88.5.375

Mozer, M. C., Pashler, H. E., & Howe, M. (2004). Using Testing to Enhance Learning: A Comparison of Two Hypotheses: (537052012-223) [Data set]. American Psychological Association. https://doi.org/10.1037/e537052012-223

Nelson, T. O., & Dunlosky, J. (1994). Norms of Paired-Associate Recall During Multitrial Learning of Swahili-English Translation Equivalents. Memory, 2(3), 325–335. https://doi.org/10.1080/09658219408258951

Paller, K. A., Kutas, M., & Mayes, A. R. (1987). Neural correlates of encoding in an incidental learning paradigm. Electroencephalography and Clinical Neurophysiology, 67(4), 360–371. https://doi.org/10.1016/0013-4694(87)90124-6

Pinheiro, J. C., & Bates, D. M. (Eds.). (2000). Linear Mixed-Effects Models: Basic Concepts and Examples. In Mixed-Effects Models in S and S-PLUS (pp. 3–56). Springer. https://doi.org/10.1007/0-387-22747-4_1

Potts, R., & Shanks, D. R. (2014). The benefit of generating errors during learning. Journal of Experimental Psychology: General, 143, 644–667. https://doi.org/10.1037/a0033194

Pyc, M. A., & Rawson, K. A. (2009). Testing the retrieval effort hypothesis: Does greater difficulty correctly recalling information lead to higher levels of memory? Journal of Memory and Language, 60(4), 437–447. https://doi.org/10.1016/j.jml.2009.01.004

R Core Team (2022). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/

Richland, L. E., Kornell, N., & Kao, L. S. (2009). The pretesting effect: Do unsuccessful retrieval attempts enhance learning? Journal of Experimental Psychology: Applied, 15, 243–257. https://doi.org/10.1037/a0016496

Rescorla, R. A., & Wagner, A. R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In Black, A. H., & Prokasy, W. F. (Eds.) Classical conditioning II: Current research and theory, (pp. 64– 99). New York: Appleton-Century-Crofts.

Roediger, H. L., & Karpicke, J. D. (2006). Test-Enhanced Learning: Taking Memory Tests Improves Long-Term Retention. Psychological Science, 17(3), 249–255. https://doi.org/10.1111/j.1467-9280.2006.01693.x

Roediger III, H. L., Putnam, A. L., & Smith, M. A. (2011). Chapter One—Ten Benefits of Testing and Their Applications to Educational Practice. In J. P. Mestre & B. H. Ross (Eds.), Psychology of Learning and Motivation (Vol. 55, pp. 1–36). Academic Press. https://doi.org/10.1016/B978-0-12-387691-1.00001-6

Rowland, C. A. (2014). The effect of testing versus restudy on retention: A meta-analytic review of the testing effect. Psychological Bulletin, 140, 1432–1463. https://doi.org/10.1037/a0037559

Seabrooke, T., Mitchell, C. J., Wills, A. J., Inkster, A. B., & Hollins, T. J. (2022). The benefits of impossible tests: Assessing the role of error-correction in the pretesting effect. Memory & Cognition, 50(2), 296–311. https://doi.org/10.3758/s13421-021-01218-6

Singmann, H., Bolker, B., Westfall, J., Aust, F., & Ben-Shachar, M. (2022). _afex: Analysis of Factorial Experiments_. R package version 1.2-0, <https://CRAN.R-project.org/package=afex>.

Singmann, H., & Kellen, D. (2019). An introduction to mixed models for experimental psychology. In New methods in cognitive psychology (pp. 4-31). Routledge.

The Power of Testing Memory: Basic Research and Implications for Educational Practice—Henry L. Roediger, Jeffrey D. Karpicke, 2006. (n.d.). Retrieved December 14, 2022, from https://journals.sagepub.com/doi/full/10.1111/j.1745-6916.2006.00012.x?casa_token=wRZOvxNkuvMAAAAA%3AstHS5IHKCatf9VSsvL1VBrxjsxwn84FUalFgPl2FLdq5mikEHfFmfarlrR629eLtgFsJa0EZHuzu

Waddill, P. J., & McDaniel, M. A. (1998). Distinctiveness effects in recall: Memory & Cognition, 26(1), 108–120. https://doi.org/10.3758/BF03211374

Wagenmakers, E.-J., Verhagen, J., Ly, A., Matzke, D., Steingroever, H., Rouder, J. N., & Morey, R. D. (2017). The Need for Bayesian Hypothesis Testing in Psychological Science. In Psychological Science Under Scrutiny (pp. 123–138). John Wiley & Sons, Ltd. https://doi.org/10.1002/9781119095910.ch8

Widrow, B., & Hoff, M. E. (1960). Adaptive switching circuits. 1960 IRE WESCON Convention Record, 96–104 Reprinted in Neurocomputing MIT Press, 1988.

Wilson, R. C., Shenhav, A., Straccia, M., & Cohen, J. D. (2019). The Eighty Five Percent Rule for optimal learning. Nature Communications, 10(1), Article 1. https://doi.org/10.1038/s41467-019-12552-4

Zheng, Y., Liu, X. L., Nishiyama, S., Ranganath, C., & O'Reilly, R. C. (2022). Correcting the hebbian mistake: Toward a fully error-driven hippocampus. PLOS Computational Biology, 18(10), e1010589. https://doi.org/10.1371/journal.pcbi.1010589

Chapter 2: Practice testing science text material enhances retention of practiced material, but non-practiced material is unaffected

Abstract

A wealth of evidence has shown that memory retrieval practice can dramatically improve retention of the information that is retrieved. Additionally, selectively retrieving *some* information can impair or improve retention of other information that was not retrieved. The effects of selective retrieval in educational texts, however, are not well understood. Some studies have shown that selective retrieval improves retention of non-practiced information if the source text was cohesive (Chan, 2009), whereas other studies have shown that, even in cohesive texts, selective retrieval can cause impairments of non-practiced information if there is competition at retrieval (Little et al., 2011). Building upon foundational reading comprehension theories that emphasize the importance of situation models (mental models of the situation described by a text; e.g., van Dijk & Kintsch, 1978), we presented an account whereby text characteristics (whether cohesively written or not), retrieval practice target characteristics (well-integrated or not), and reader characteristics (prior knowledge and reading ability) all contribute to the sequelae of selective retrieval for non-practiced information. Text and reader characteristics contribute to the reader's ability to form a coherent mental model of the text, and retrieval practice target characteristics dictate the extent that information is reactivated during retrieval practice. Main ideas of text are more likely to be highly integrated into the situation model of the text, whereas peripheral, supporting ideas, are not, suggesting that retrieval practice of main ideas is most likely to benefit retention of non-practiced information. Here, using cohesive science texts, participants practiced retrieval of main ideas in one group, practiced retrieval of

peripheral ideas in a second group, or were excused until the final tests in the control group. On final tests, we predicted that both retrieval practice conditions would enhance retention of previously practiced information. Critically, for non-practiced information, we predicted that main idea retrieval practice would lead to retention benefits, whereas peripheral idea retrieval practice would lead to retention impairments, relative to the control condition. Results from two experiments using distinct texts revealed robust effects of retrieval practice for practiced information, but retrieval practice had modest and inconsistent effects for non-practiced information. Individual differences in prior knowledge and reading ability had more robust effects on retention of non-practiced information. We conclude by discussing how our results fit with the situation model account and the implications for educational practices.

Introduction

Imagine that you have just read an article covering the basics of virus reproduction and transmission and then you recall a specific selection in a casual conversation. You might recall one of the main points, such as that viruses can only reproduce inside a living cell, or a tangential, yet interesting, detail, such as that one of the best-studied viruses, known as the bacteriophage, infects bacteria. In this scenario, your memory would likely perform satisfyingly, providing you with the information you searched for without overwhelming you with a flood of everything you learned all at once. Now imagine the next day, you are attempting to explain to a family member how vaccines protect against viral illness, and you need all the details at your disposal to make a cohesive argument. Intuitively, it might seem as if your memory would be unchanged by what you recalled the day before, but research on memory has indicated otherwise.

A wealth of evidence has shown that the act of memory retrieval can dramatically improve retention of the information that is retrieved. Additionally, selectively retrieving *some* information can impair or improve retention of other information that was not retrieved. We engage in selective retrieval constantly, and in educational contexts, selective retrieval is integral to content reviews and quizzes. Thus, it is of critical importance to understand whether or how selective retrieval impacts retention of information that is not retrieved. Here, we explored selective retrieval using educationally relevant retrieval practice and materials that are well-motivated to lead to a facilitative effect. Lastly, we introduce a framework for understanding the sequelae of selective retrieval in prose text grounded in reading comprehension theory.

Retrieval practice–the act of attempting to retrieve information given a cue–has proven to be one of the most potent and time efficient memory enhancement techniques when compared to restudying. The retention benefits of retrieval practice have borne out across diverse

experimental conditions, stimuli, and educational settings (Roediger & Karpicke, 2006; Rowland, 2014, Adesope et al., 2017; Agarwal et al., 2021). This robust body of evidence has informed researchers' recommendations that educators prioritize retrieval practice (e.g., Roediger & Karpicke, 2006; Roediger et al., 2011). Although Hinze et al. (2013) provided support for a constructive retrieval hypothesis whereby open-ended retrieval practice that engages constructive practices leads to greater retention and comprehension than rote retrieval practice, open-ended retrieval practice does not permit investigation of practice effects on non-practiced material. Retrieval practice undoubtedly benefits the retention of target content, but the effects of selective retrieval on related, non-practiced material are more complex.

A great deal of research has shown that, under certain circumstances, selective retrieval can impair retention of related, non-practiced information, a phenomenon known as "retrieval-induced forgetting" (RIF; Anderson, 2003; Anderson et al., 1994). For instance, Anderson et al. (1994) had participants study category-exemplar pairs with each category having two targets (metal-silver; metal-iron; fruit-apple; fruit-banana). During the practice phase, participants practiced retrieval for one cue-target pair from some of the categories with letter support (metal-si___). In the final phase, participants completed a cued-recall criterial test consisting of three trial types: target retrieval practiced pairs (RP+; metal-silver), the related, non-practiced pairs (RP-; metal-iron), and the unrelated, non-practiced pairs (NRP; fruit-apple, fruit-banana). Consistent with studies of the testing effect, participants showed better retention for the RP+ pairs than both the RP- and NRP pairs. In contrast, retention of RP- pairs was worse than NRP pairs. Theoretical underpinnings of RIF have long been debated, but what is agreed upon is that there is competition between the RP+ target and the RP- target. Some have argued that the mechanism that brings about competition is inhibition of the RP- items during retrieval practice,

as Anderson (2003) suggests, and others have argued that competition arises from interference during the final test (Raaijmakers & Jakab, 2013). Anderson et al.'s results, and the numerous replications of the RIF effect (Murayama et al., 2014), demonstrated that selective retrieval can adversely impact retention of related, non-practiced material.

In other situations, selective retrieval has been shown to enhance retrieval of related, non-practiced information, a finding known as "retrieval-induced facilitation" (RIFA; Chan et al., 2006). The RIFA effect is observed when the retention benefits of selective retrieval "spill over" to related, non-practiced material (RP- > NRP), instead of impairing its retention. Chan et al. (2006) observed a RIFA effect for the first time in a cued-recall final test. In Experiment 1, participants read a prose passage about Toucans for 25 minutes, then one group practiced retrieval for one half of the cued-recall question bank, a restudy group reviewed the same questions and their answers, and a control group was dismissed. Twenty-four hours later, participants from all groups completed the final test containing the full set of cued-recall questions. Analysis of related, non-practiced questions was of the greatest interest, which revealed greater performance for the retrieval practice group than both the restudy and the control group, hence RIFA. Chan et al. replicated the RIFA effect in Experiments 2 and 3, concluding that they could make a relatively unreserved recommendation for frequent classroom testing (Chan et al. 2006, p. 566). The implications of this result were related to known boundary conditions in RIF, particularly integration (Anderson, 2003; Anderson & McCulloch, 1999) and a one day interval between retrieval practice and the final test (MacLeod & Macrae, 2001). Anderson and McCulloch (1999) found that a simple instruction to integrate stimuli during encoding was sufficient to significantly reduce RIF in a paradigm that was otherwise identical to Anderson et al. (1994). The materials Chan et al. (2006) used may have been automatically

"integrated" during encoding due to typical properties of cohesive prose texts, therefore the RP- material was protected from the competition that would have caused RIF otherwise.

Contemporary theories of comprehension generally assume that the goal of text comprehension is to form locally coherent (within text) and globally coherent (within prior knowledge) mental models (A. C. Graesser et al., 1994; McNamara & Magliano, 2009). When coherence breaks down, readers attempt to repair it by making inferences. Inferencing is the process of connecting information that is currently being processed to information that is not currently being processed (e.g., from the current sentence to an earlier sentence). Inferences either refer back to earlier in the discourse (bridging), or out to prior knowledge (elaborations; McNamara and Magliano, 2009). Coherence of the explicit text, known as *cohesion*, consists of characteristics of the text that play a role in helping the reader mentally connect the ideas in the text (Graesser et al., 2003). Cohesion of the text therefore influences comprehension and memory (e.g., Cohn-Sheehy et al., 2022; McNamara et al., 1996). Accordingly, Chan et al.'s (2006) texts may have implicitly encouraged integration by being written cohesively.

In follow-up work to Chan et al. (2006), Chan (2009) sought to identify the boundary conditions between RIF and RIFA. Chan manipulated the delay between retrieval practice and the final test (20 minutes or 24 hours) and integration. In the high integration condition, the texts were presented in a normal, cohesive order similar to an educational expository text and participants were instructed to integrate ideas as they read. In the low integration condition, participants were presented with the same sentences in a random, incohesive order and were instructed to memorize a list of facts. In Experiment 1, for the 20-minute delayed test groups, RIF was observed for the low integration condition, but there was a null effect for the high integration condition. For the 24-hour dealayed test groups, RIFA was observed for the high

integration condition, and there was a null effect for the low integration condition. The presence of RIF only with a short delay and low integration, and the RIFA effect only with a long delay and high integration suggests that integration protects against RIF, and delay distinctly benefits related, non-practiced material. Experiment 2 replicated Experiment 1's results using stimuli that resembled the simplicity of classic category-exemplar stimuli (e.g., "the fork is in the nursery", "the painting is in the nursery"), but did not rely on semantic knowledge. Chan (2009) concluded that text cohesion protects against RIF in educationally relevant materials and put forward an account of selective retrieval in prose texts inspired by theories of reading comprehension. We will present Chan's theory after reviewing two further instances of RIFA.

Chan found that integration plays a critical role not only in protecting against RIF, but in producing a RIFA effect as well. In the next example of a RIFA effect, Jonker et al. (2018) manipulated integration in a design that used visual stimuli, which provided evidence that the RIFA effect is not limited to text materials. Jonker et al. (2018) were interested in whether the testing effect and/or RIFA resulted from the reactivation of episodic context, operationalized as visual scenes that were presented for two consecutive trials. In order to explore this possibility, they developed a paradigm that resembled the classic RIF paradigm using visual stimuli. On day one, participants encoded scene-object pairs such as a picture of a campground and a picture of an avocado. Critically, each scene was associated with two unique objects, the same scene was used for two consecutive trials, and participants were instructed to imagine both objects in the same scene for the specific purpose of encouraging integration. After encoding, participants practiced retrieval for some scene-object pairs, and restudied other scene-object pairs. Critically, there was re-exposure to only one of the two objects associated with each scene. In Experiment 1, there were either one or three repetitions of re-exposure. The next day, participants completed

the final test containing the full stimulus set. Experiment 1 revealed a RIFA effect in that the objects that shared scenes with the objects practiced three times were better recalled than the objects that shared scenes with restudied objects. This result emphasized the importance of repeated retrieval to RIFA because the effect was not observed for the single repetition group. Moreover, the results suggest that the effects of integration in protecting against competition and producing RIFA generalize beyond text materials to objects in visual scenes.

Jonker et al.'s (2018) results suggested that the shared features that enable RIFA can be temporal and visual, in addition to the conceptual and spatial overlap of practiced and non-practiced materials in the prose materials presented by Chan et al. (2006) and Chan (2006). In view of these results, Liu and Ranganath (2021) were interested in further investigating the features that can produce a RIFA effect. Liu and Ranganath (2021) also used visual scenes, but, instead of pairing scenes with visual objects, the scenes were paired with words. The two words, "pairmates", were either semantically related to the scene or unrelated, and the pairmates of a scene were either temporally close or far apart. The third factor they manipulated was whether or not there was a delay that included sleep. After encoding, half of the pairmates were practiced over two repetitions. Across three experiments, their results revealed that selective retrieval produced a RIFA effect in temporally close RP- words, and a RIF effect in RP- words that were temporally far and semantically unrelated. Intriguingly, sleep mediated the boundary between RIF and RIFA, such that a RIF effect was observed for temporally far and related RP- words without sleep, but a RIFA effect was observed for the same condition when participants slept before the final test. Taken together, these results suggest that multiple stimulus dimensions influence integration independently. With sleep, semantic relatedness can overcome competition caused by temporal distance. These results underscore the potential for retrieval practice to

provide retention benefits for target and non-practiced material, particularly in situations that students are likely to encounter, such as when tested the day after studying.

In studies using prose texts and relatively simple stimuli, a number of factors have been associated with observing a RIFA effect–cohesive text, instructions to integrate, close temporal and semantic distance, and sleep–raising the prospect that they involve the same underlying mechanism. One possibility put forward by Chan (2009) was rooted in the notion that readers form "situation models" as a product of comprehension (van Dijk & Kintsch, 1983). Situation models are mental representations of the situation described by the text and integrated with prior knowledge. They encompass connections among ideas in a text and connections to prior knowledge. To advance his theory, Chan drew a comparison between his experiments and an interesting finding from Radvansky and Zacks (1991) involving situation models. Radvansky and Zacks (1991) created three conditions from verbal location-object pairings, similar to Chan (2009)'s Experiment 2 stimuli: in Condition 1, one object appeared in one location; in Condition 2, multiple objects appeared in one location; in Condition 3, one object appeared in multiple locations. A fan effect was observed in Condition 3, but not in Condition 2. Radvansky and Zacks (1991) concluded that objects that share the same location can be integrated into the same situation model. Paralleling this result, Chan (2009) suggested that the high integration condition encouraged participants to form fewer situation models than in the low integration condition, thereby reducing the overall level of competition between situation models. A reduction in competition on its own would not lead to a RIFA effect, therefore Chan (2009) further clarified his account, asserting that selective retrieval strengthens information in the same situation model as retrieved information. Finally, the delay between selective retrieval and the final test is assumed to provide an additive effect to integration against competition.

If we view Jonker et al.'s (2018) results through the lens of Chan's theory, a more general account emerges for when RIFA effects occur. Jonker et al. encouraged participants to imagine the target objects in the same scene cued by the scene image, and trials that shared a scene were temporally contiguous. In a sense, they were asked to comprehend the scene and object pairings, which would activate relevant prior knowledge and connect the objects into a coherent situation model. Indeed, although comprehension theory has largely been developed in the context of text processing, many theorists recognized that comprehension is a general cognitive process (Gernsbacher et al., 1990; Kintsch, 1998). In selective retrieval practice, we argue that cueing with the scene and target word's first letter reactivated and strengthened the whole representation, thereby strengthening the non-practiced object. Liu and Ranganath (2021) used a similar paradigm and showed that temporally close and semantically related pairwords could demonstrate a RIFA effect. Intriguingly, temporally far, related pairwords showed a RIFA effect too – but only with sleep. We argue that these factors increased the likelihood that pairwords would be integrated into a single situation model. To summarize, RIFA effects are more likely to occur when conditions allow strong integration between target stimuli by way of a situation model representation.

We now turn our focus to our primary interest, selective retrieval in educational texts. The studies displaying RIFA effects so far suggested that selective retrieval in educational prose texts is likely to produce RIFA effects. Educational prose texts can be expected to be well-organized and cohesive, which entails that the most related ideas are presented in nearby sentences. Information that is less related is presented in separate paragraphs, and a cohesively written text will make explicit the connections between less related ideas. Notably, however,

RIFA is not ubiquitous in selective retrieval studies of cohesive prose text and sometimes RIF occurs.

Carroll et al. (2007) were interested in the effects of selective retrieval when the materials were educationally relevant texts. In Experiment 1, high and low prior knowledge participants studied two abnormal psychology passages then practiced retrieval on half of the cued-recall questions for one passage. After 15 minutes in the immediate test group, or 24 hours in the delayed test group, participants completed the final criterial test containing the full set of cued-recall questions. The critical outcome was the difference in recall between RP- and NRP questions. Only the novice group demonstrated RIF at the immediate test, suggesting that the prior knowledge of the experts afforded protection against RIF, perhaps by facilitating integration and organization of the material. Although a RIFA effect was not observed, the 24-hour delay evidently provided protection against RIF for novices that was not sufficient at the 15 minute delayed test. In Experiment 2, Carroll et al. sought to test the integration hypothesis further by manipulating the text passage cohesion by randomly presenting its sentences, or presenting them in the normal, cohesive order. An expert group and a 24 hour delay group were not included in Experiment 2. Carroll et al. predicted that the random order text would be more susceptible to RIF by disrupting the integration naturally afforded by a cohesive text. Furthermore, they used three different final test types (multiple choice, short answer, and essay) to increase the ecological validity of their results. They observed RIF for both the high and low cohesion text conditions, and RIF was present for short answer and essay tests but not for multiple choice. These results suggest that cohesive text alone cannot protect against RIF at an immediate test, but that prior knowledge may protect against RIF. However, at a delayed test, RIF was not observed. A subtle but potentially important design feature, however, is that there

was likely competition between the two texts given that they are within the same subject area. Using two texts from different domains would likely reduce competition, but a between-participants design would eliminate this source of competition completely. To summarize, the presence of RIF, even in cohesive texts, suggests that cohesion alone might not encourage sufficient integration to counteract the competition introduced by selective retrieval.

Little et al. (2011) were interested in why cohesive text appeared to protect against RIF in some studies (Chan et al. 2006; Chan, 2009), but did not in other studies (Carroll et al. 2007). The principal difference appeared to be in the text materials, which Chan et al. carefully composed from several source articles and included "facilitative questions", whereas Carroll et al.'s passages were essentially verbatim excerpts from a textbook and randomly sampled questions. Chan et al. created two sets of questions because the questions in one set were likely to cause activation of the target content in the other set. For example, one sentence in the Toucans passage contained two questions, one from each set (Its tongue is like a *feather* which is used to catch food and *flick it down its throat*; italicized text indicates correct answers to fill-in-the-blank questions from each set). Not all of the question pairs were drawn from the same sentence, but this example serves to show that the structure of the texts bore a greater resemblance to the place and object propositions used in Chan (2009) Experiment 2, than to a typical educational text passage. For these reasons, Little et al. recrafted Chan's questions to eliminate the facilitative relationship between questions. Little et al. used a 15-minute delayed test after retrieval practice. In Experiment 1, Little et al. observed no significant differences between RP- and NRP items, replicating Chan et al.'s (2009) null effect at the immediate test, but without facilitative questions. The main question for Little et al. was why RIF was not observed under these conditions. In Experiment 2, the text sentences were randomized to

decrease the likelihood of spontaneous integration, in the expectation that now RIF would be observed. Indeed, RIF occurred, replicating Chan's (2009) result in the low integration, short delay condition. Little et al. were perplexed at why Carroll et al. observed RIF with cohesive text, but Chan et al. and Little et al. Experiments 1 and 2 did not. Given these inconsistent results, in Experiments 3 and 4, Little et al. chose to introduce greater retrieval competition while still using cohesive, educationally relevant text passages. Instead of two distinct text passages, they created six cohesive texts about distinct locations that were structured by the following three headings: geography, climate, and people. In the retrieval practice group, participants were instructed to write short essays about the geography, climate, and people for the three locations with which they were prompted, and the NRP group completed a distractor task. In the criterial essay test, both groups were prompted to write about the same three non-practiced locations. Therefore, in this design, RP+ was not probed at the final test and the RP- content is related only in that it shares the headings of geography, climate, and people; NRP was the recall of a control group of participants. In accordance with Little et al.'s predictions, Experiments 3 and 4 produced RIF despite using cohesive text, providing evidence that retrieval competition is critical to producing RIF in prose texts.

To sum up, studies examining selective retrieval of complex prose materials have revealed conflicting results. In some circumstances, selective retrieval practice was associated with RIF or a null effect, and in other circumstances it was associated with RIFA. Accordingly, two countervailing claims have been made: situation model integration facilitated by cohesive text causes RIFA, and retrieval competition, even in cohesive texts, produces RIF. In our assessment, there are at least two factors that might contribute to this variance: characteristics of the text and individual differences in the ability to integrate the text into situation models. We

hypothesized that both of these factors should influence whether selective retrieval causes RIF or RIFA. Moreover, prior studies of retrieval in complex materials considered text cohesion as a binary variable—low cohesion texts were not prose texts at all, they were random sequences of sentences–but this approach ignores the complex connections that occur between ideas in the text and in the reader's situation models of the text. In Kintsch's (1998) Construction-Integration model of comprehension, overlap between sentence arguments provides the principal means of connecting ideas, but inferences make connections, too. The construction phase involves activating concepts and ideas bottom-up based on available retrieval cues, and in the integration phase, activation first spreads through the activated concepts, then concepts with fewer connections are down-weighted, ultimately leaving those activated which have the most connections to other concepts. One implication is that the main idea sentence of a text should have the most connections with the whole text in a reader's situation model, whereas a peripheral, supporting sentence might only have one or two connections, or none at all. Therefore, it is important to consider the relationship between retrieval practice targets and non-practiced information. For instance, a main idea will be more likely to be well integrated into a mental representation of the text, so when its content is cued for selective retrieval, the integration should protect the related, non-practiced material against competition. In contrast, a peripheral idea will be more likely to be poorly integrated into the mental representation of the text, so selective retrieval of peripheral ideas should increase competition with other ideas in the text. Accordingly, we might expect selective retrieval of main ideas to cause RIFA and selective retrieval of peripheral information to cause RIF. To our knowledge, no studies have investigated whether differently integrated retrieval prompts cause RIF or RIFA.

If reading comprehension theory is taken into consideration, then the question of how to probe selective retrieval in prose text becomes clearer. First, it's important to use text materials that are ecologically valid, so we used short introductory science passages on viruses and the endocrine system. Second, the cues for selective retrieval should be selected based on criteria that can be replicated in new texts, so we used sentences extracted from the text material, as opposed to bespoke cued-recall questions. Finally, the criterial test should index the full extent of participants' retention, so we used a free recall criterial test scored by idea units in addition to a multiple-choice test that indexed detail and inference knowledge.

The predicted outcomes of main idea and peripheral idea retrieval practice are predicated on the assumption that readers form a mental model that captures the relational structure of the text. Of course, this assumption will not always hold true. In addition to text cohesion, individual differences in reading ability and prior knowledge also contribute to forming a coherent mental model of a text (e.g., O'Reilly & McNamara, 2007). When coherence breaks down, skilled readers are better equipped to make bridging inferences within the text to maintain coherent situation models (e.g., Voss & Silfies, 1996). Therefore, we predicted that these individual difference variables would be associated with recall of non-practiced idea units and performance on the multiple-choice test. Moreover, given the situation model theory described here, we predicted that good readers and high knowledge participants would be more likely to demonstrate RIFA following selective retrieval. In contrast, poor readers and novices might find it more difficult to build coherent situation models from the text (or might build too many cf., Gernsbacher, 1990). Therefore, we predicted they will be more likely to demonstrate RIF following selective retrieval.

Here, we examined the extent to which retrieval targets and individual differences in reader skills and knowledge determine the effects of selective retrieval on retention of non-practiced information. In Experiment 1, participants studied a science text and were tested repeatedly on target information from either main idea or peripheral sentences in the text. In a final free recall test, subjects were asked to recall the entire text, then complete a multiple-choice test. We predicted that recall of both targets and previously non-practiced idea units would be higher for participants who were tested on main ideas than for participants who were in the control group. For participants who were tested on peripheral sentences, we predicted that recall would be higher for target idea units, but worse for non-practiced idea units, when compared to the control group. Furthermore, given the critical role of reading ability and prior knowledge to forming a coherent mental model of a text, we predicted that these individual difference measures would moderate retention and explored whether they moderated any effects of selective retrieval.

Methods

Participants

In Experiment 1, 202 participants recruited through Amazon Mechanical Turk completed all tasks and were paid $9. Thirty-two participants were excluded from all analyses due to failing data quality checks, leaving 170 participants in the final sample. In Experiment 2, 223 participants recruited through Amazon Mechanical Turk completed all tasks and were paid $9. Thirty-three participants were excluded from all analyses due to failing data quality checks, leaving 190 participants in the final sample. Based on the effect size of the RIFA effect observed in Chan (2009), we conducted a power analysis in GPower (Faul et al., 2007) for a difference in

means between unmatched pairs with power set at 0.80 and alpha = 0.05, which indicated 57 participants per group were required to detect an effect size of Cohen's $d = 0.53$. In the final sample of Experiment 1, there were 61 participants in RPm, 54 in RPp, and 55 in NRP. In Experiment 2, there were 62 participants in RPm, 62 in RPp, and 66 in NRP. This study was approved by the UC Davis Institutional Review Board.

Materials

Both of the text passages were previously used in Hinze et al. (2013) with minor modifications. In Experiment 1, participants read a 373-word passage about how viruses work. In

author	Cohesion-Deep	Cohesion-Referential	Flesch-Kincaid
Chan (2009)	0.07	−0.41	1.94
Little et al., (2011)	−0.80	−1.15	9.33
Reilly et al., (in press)	0.39	1.41	5.61

Experiment 2, participants read a 456-word passage about the endocrine system and the fight-or-flight response. Due to the emphasis placed upon text cohesion, we used natural language processing tool called Cohmetrix (A. C. Graesser et al., 2004) to compute two cohesion metrics as well as the Flesch-Kincaid grade level of our texts and the texts used by Chan (2009) and Little et al. (2011). As can be seen in Table 1, the texts in the current study scored much higher on average on z-scale Deep Cohesion and Referential Cohesion.

Table 1. Comparisons of text cohesion and text difficulty derived from *Cohmetrix* (Graesser et al., 2004). Text cohesion metrics are on a z-scale.

Design

Participants read an educational text passage then were randomly assigned to one of three groups in a between-participants design. There were two retrieval practice groups and a control group. For each paragraph, a main idea sentence and a peripheral, supporting sentence were identified. Participants in the main idea retrieval practice group (RPm) completed fill-in-the-blank retrieval practice with feedback for each of the four main idea sentences of the passage, one from each paragraph. Likewise, in the peripheral idea group (RPp), the same task was completed for the peripheral idea sentences. Twenty-four hours later, participants from all groups completed free recall and multiple choice for the studied passage, answered prior knowledge questions, and completed the reading ability measure.

Experiment 1 Procedure

Part 1. After completing the informed consent, participants read the task instructions. Participants were instructed to carefully read the text passage and were told that their memory and understanding would be tested. The text passage was presented one sentence at a time and was self-paced. After each paragraph's sentences were presented, the whole paragraph was shown again. In this way, participants were guided to read each individual sentence, and also to integrate sentences within each paragraph, or look back at a sentence. After reading the passage, participants in the control group (NRP) were excused. Participants in the retrieval practice group immediately moved on to the retrieval practice instructions.

Part 2. Participants in the retrieval practice groups completed selective retrieval practice immediately after reading the text. In the RPm group, participants completed cued-recall for each

main idea sentence in each paragraph of the text. In the RPp group, participants completed cued-recall for each peripheral sentence in each paragraph of the text. For each practiced sentence, five content words were not shown. Participants were instructed to provide responses for each of the five missing words. After three incorrect responses to the first missing word, correct answer feedback was shown, and this was repeated for each of the five words. Only one response could be scored per trial, therefore participants received feedback in sequential order, rather than in the order that their correct responses were made. Therefore, for each sentence, the minimum number of retrieval attempts was five, and the maximum was fifteen.

Part 3. Twenty-four hours after Part 1 began, participants completed the self-paced final tests and individual difference measures. First participants completed the cued recall test. They were instructed to recall as much as they could from the text by entering in one complete thought at a time. In this way participants were encouraged to free recall, rather than composing and editing an essay response. After the recall test, participants completed the multiple-choice test. The order of the test questions was randomized and participants completed one question at a time. Next, participants answered a questionnaire which included questions about their effort and use of outside sources as data quality checks, and we asked how familiar they were with the content of the text passage prior to the experiment using a slider from zero to one hundred to be used as the prior knowledge measure. Finally, participants completed the Gates-MacGintie Reading Test adapted for online administration, which tested their comprehension and vocabulary knowledge (MacGinitie et al., 1989)

Experiment 1 Analysis Methods

Recall Scoring

We were interested in the retention of the complete text passage content after 24 hours. Therefore, we scored the recall responses according to the idea units (conceptually meaningful parts of sentences) that they included. In Experiment 1, the viruses text was parsed into 48 idea units. For each recall, as opposed to a purely aggregate approach, idea units were awarded for each individual response statement or sentence, in the order that it was recalled. An idea unit was awarded if the statement captured most of its content. Responses did not need to be verbatim, but they were required to be specific enough to assign them to a particular idea unit. Gist or summary statements were not awarded idea units. Responses stemming from knowledge outside the text, whether correct or not, were not assigned to idea units. Two raters blinded to condition scored the first 100 subjects' recalls, with an inter-rater reliability of .72. After establishing agreement, one rater scored the remaining recalls. The idea units that corresponded to retrieval practiced sentences were of particular interest to the present experiments, but the same scoring regime was identical for them as for non-practiced idea units. In the viruses text, eight idea units corresponded to the four main idea sentences and nine idea units corresponded to the four peripheral idea sentences. In Experiment 2, the endocrine text was parsed into 42 idea units. One rater scored all of the endocrine text recalls. In the endocrine text, seven idea units corresponded to the four main idea sentences and seven idea units corresponded to the four peripheral idea sentences.

Recall Analysis

We were interested in how recall of the three types of idea units was affected by participant group, reading ability, prior knowledge, and their interactions, while controlling for variance due to individual subjects and individual idea units. We created three generalized mixed-effects models (Jaeger, 2008) with a binomial distribution and a logistic link function for recall of the three types of idea units: main idea units, peripheral idea units, and non-practiced idea units. The idea units binned into these analyses were the same for all participant groups (i.e., there was no conditionalization of idea units). All three mixed-effects models included the same fixed and random effects. The model estimated fixed effects for participant group, reading ability, prior knowledge, and up to three-way interactions of these predictors, and random intercepts for participant identity and idea unit. The variables for participant group, id, and idea unit were each coded with sum-to-zero contrasts in the model. The models were fit using the afex library's function, 'mixed', (Singmann et al., 2022) which used lme4's 'glmer' function (Bates et al., 2015).

We used a parametric bootstrap, type III sum of squares method for predictor inference which was recommended for experiments similar to the present ones (Singmann & Kellen, 2019). There are difficulties in inference for generalized mixed effects models due to the inability to estimate denominator degrees of freedom, and inflated type 1 errors are associated with likelihood ratio tests, particularly when random factors have fewer than 40 levels (Pinheiro & Bates, 2000). Therefore, predictor inference was carried out with the pbkrtest library's 'PBmodcomp' function. In a similar way to a likelihood-ratio test procedure, PBmodcomp compares a reduced model to the full model in order to make inferences about a left-out predictor. The PBmodcomp parametric bootstrap procedure simulates datasets from the reduced

model, then fits both the reduced and the full model to each dataset. The parametric bootstrap p-value corresponds to the percentage of simulated likelihood-ratio values that are larger than the observed likelihood-ratio value (Halekoh & Højsgaard, 2014). We used R (version 4.2.2, R Core Team, 2022) to compute all analyses.

We conducted Bayes Factor analyses to investigate whether our data provided evidence for the null hypothesis or were, instead, merely insensitive to our manipulation. Specifically, we tested the null hypothesis against two theoretical priors given the data we observed. The null hypothesis was that there would be no impact on retention of related material due to retrieval practice, which was represented as a single point at zero. Following Dienes's (2014) guidelines and anchored by Chan's (2009) observation of both RIF and RIFA effects equal to 9%, the RIFA hypothesis was represented by a half-normal distribution with a minimum of zero and a standard deviation of 9, and the RIF hypothesis was represented by the negative of the same distribution. The choice to use half-normal distributions for priors instead of a normal distribution stacked the deck in favor of observing evidence for the alternative hypothesis, therefore any evidence for the null hypothesis could be interpreted as particularly strong. The data model was represented by a normal distribution with a mean of the observed difference in non-practiced idea unit retention between RPm and NRP, and between RPp and NRP, and a standard deviation of one half of the mean difference (Dienes, 2014). Bayes Factors were computed by first weighting the data models by the three prior distributions (RIFA, RIF, and null), then dividing the integral of the RIFA weighted distribution by the integral of the null distribution, and likewise, dividing the integral of the RIF weighted distribution by the integral of the null distribution. Bayes Factor values range from 0 to infinity, whereby 1 indicates equal likelihood of the null and alternative distribution, and values between .3 and 3 are perceived to indicate insensitivity to detecting

differences. We used the Bayesplay library to compute Bayes Factors and Wagenmakers et al. (2017) interpretations of Bayes Factor values.

Multiple Choice Analysis

The final multiple-choice test was analyzed with a generalized mixed-effects model in the same manner as the recall models, including the predictors participant group, reading ability, and prior knowledge, with the addition of a predictor for question type, which included levels for "detail" and "inference" questions. Detail questions could be answered correctly based on the content of a single isolated sentence, whereas inference questions required making a connection between separate sentences. Inference was based on the same parametric bootstrap procedure described in the recall analysis methods section above.

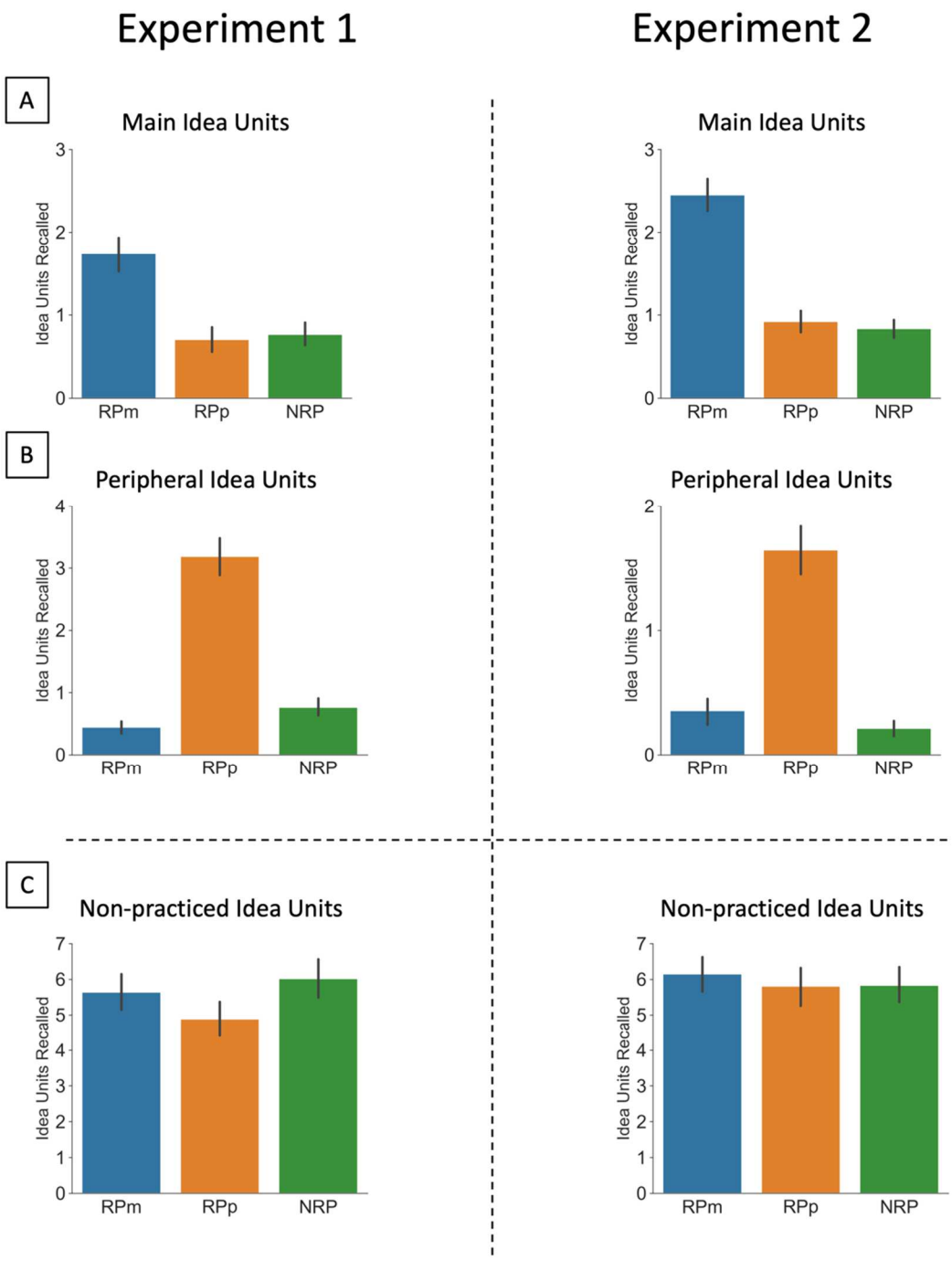

Figure 2.1. Recall Test Performance. Mean idea units recalled on the final free recall test for each type of idea unit for Experiment 1 (viruses text) and Experiment 2 (endocrine text). Main idea units were practiced by the RPm group only. Peripheral idea units were practiced by the RPp group only. The NRP group did not practice any idea units. Panels A and B show the "testing effect" in that retention of main ideas and peripheral ideas were greatest for the RPm and RPp groups, respectively. Panel C depicts recall of non-practiced idea units by all three condition groups. Error bars indicate standard error of the mean.

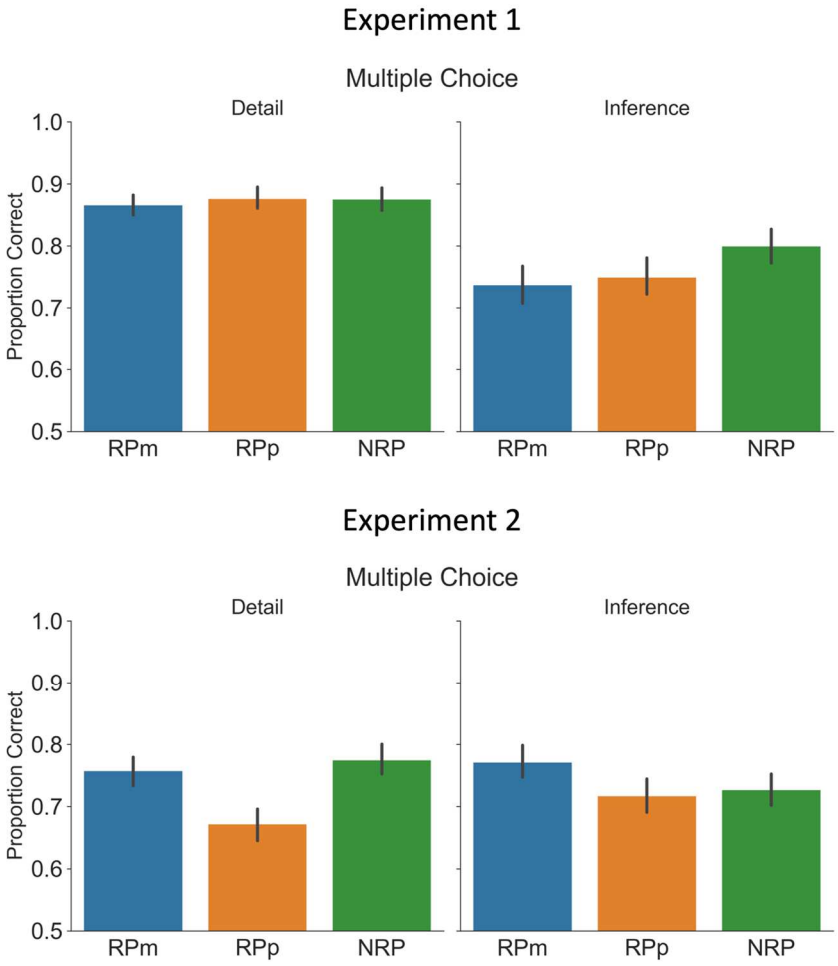

Figure 2.2. Multiple-choice Test Performance. Mean proportion correct on the multiple-choice final test. Error bars represent standard error of the mean.

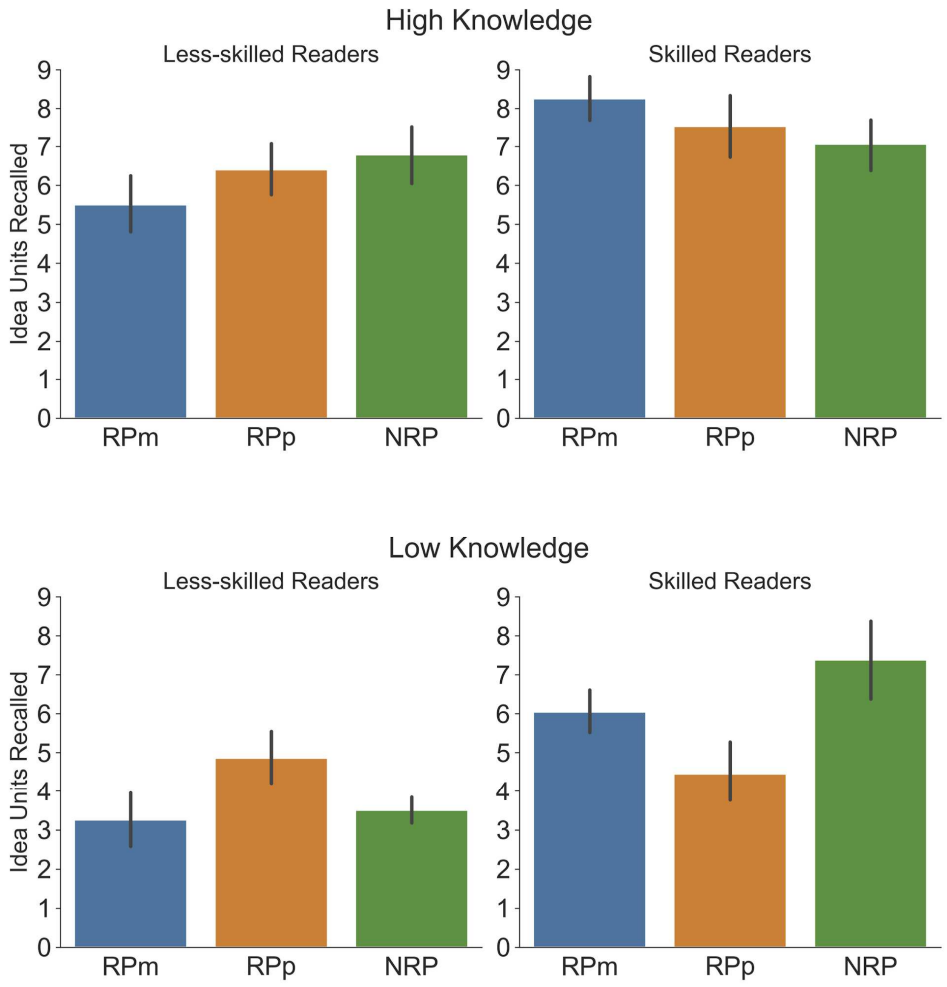

Figure 2.3. Effects of reading skill and domain knowledge on recall in Experiment 2. Bootstrap analyses of non-practiced idea units in the endocrine text (Experiment 2) revealed a significant three-way interaction between participant group, reading ability, and prior knowledge. For visualization purposes, participants were divided with a median split according to prior knowledge and reading ability. Error bars represent standard error of the mean of each cell.

Experiment 1 Results

Control analyses: retrieval practice conditions boosted memory for practiced material

Before proceeding to examine effects of retrieval practice on memory performance, we ran analyses to rule out variables that could potentially confound between-group comparisons:

the amount of time spent reading the text during part 1, individual differences in reading ability, and prior knowledge of the topic of the text. One-way ANOVAs revealed no significant between-group differences in these variables (all F's <1.5, all p's > .27).

Our next analyses focused on the effects of retrieval practice on initial learning. Comparisons between the two retrieval practice groups revealed that, on average, those in the RPp group required more practice attempts than those in the RPm group in order to achieve criterion performance on the tested idea units (t (113) = 2.72, p = .007), which is consistent with the idea that main ideas are retained more accurately than peripheral ideas. Despite this difference, as we will describe below, both participant groups demonstrated robust testing effects.

Our hypotheses focused on the effects of retrieval practice on retention of tested and non-practiced idea units. To verify the efficacy of our testing manipulation, we first examined retention of idea units that were practiced in either the RPm or RPp groups. Participants in the RPm group showed better retention of the practiced main idea units than did participants in the RPp and NRP groups, who were not previously tested on these sections (Figure 1A). As described in the recall analysis section, we conducted separate mixed-effects models for each type of idea unit, and each model estimated fixed effects of participant group, reading ability, and prior knowledge, and controlled for participant identity and idea unit with random intercepts. Predictor inference was computed using a parametric bootstrap, type III sum of squares method, also described in the analysis methods. For main idea units, the analysis revealed a significant effect of participant group (X^2 = 23.49, p < .001), and follow-up comparisons revealed significantly greater retention in the RPm group compared to the NRP group (z = 3.625, p < .001), but the same comparison for the RPp group was not significant (z = .97, p = .33). These

results confirmed that selective retrieval of main idea units in the RPm group results in a testing effect.

Conversely, participants in the RPp group showed better retention of the practiced peripheral idea units than did participants in the RPm and NRP groups, who were not previously tested on these sections (Figure 1B). The mixed-effects model for peripheral idea units revealed a significant effect of participant group ($X^2 = 84.59$, $p < .001$). Follow up comparisons revealed a significant effect of RPp compared to NRP ($z = 7.17$, $p < .001$), but the same comparison for the RPm group was not significant ($z = 1.51$, $p = .13$). These findings indicate that our testing manipulation enhanced retention of practiced material, consistent with other studies of the testing effect.

Recall of non-practiced idea units: No evidence for RIFA, weak evidence for RIF in the RPp condition, and a positive effect of reading ability

We next turned to retention of idea units that were not previously tested in any of the three groups (Figure 1C), for which we predicted a RIFA effect for the RPm group and a RIF effect for the RPp group. We also predicted that reading ability and prior knowledge would moderate retention of non-practiced units. Contrary to our predictions, the parametric bootstrap predictor inference did not reveal a significant effect of participant group ($X^2 = 6.20$, $p = .082$), or prior knowledge ($X^2 = 2.81$, $p = .116$); however, the model revealed a significant effect of reading ability ($B = .282$, $X^2 = 11.38$, $p = .003$), indicating that reading ability was positively related to non-practiced idea unit recall.

Given the surprising null effect of participant group on retention, we conducted Bayes Factor analyses to investigate whether our data provided evidence for the null hypothesis or were, instead, merely insensitive to our manipulation (see Methods for details). The difference in

retention between the RPm condition and the NRP condition yielded very strong evidence for the null hypothesis over the RIFA hypothesis (BF = .012) and insensitivity between the RIF and the null hypothesis (BF = .516). The difference in retention between the RPp condition and the NRP condition yielded strong evidence for the null hypothesis over the RIFA hypothesis (BF = .087) and moderate evidence for the RIF hypothesis (BF = 3.192). Taken together, these results indicate that selective retrieval did not produce our hypothesized RIFA effect for non-practiced units. Conversely, there was moderate evidence for a RIF effect for the RPp group.

Multiple Choice: No significant main effect of participant group, RPm benefitted low knowledge participants, positive effect of reading ability

In the multiple-choice test, the inference questions were typical of a comprehension test, requiring synthesis of information drawn from more than one source sentence, whereas the detail questions required retention of facts. On multiple-choice detail questions, our predictions were the same as for non-practiced idea units: we predicted better retention in the RPm than NRP condition and NRP than RPp condition. On inference questions, we predicted that prior knowledge and reading ability would be positively correlated with performance.

We found no significant between-groups differences on multiple-choice test accuracy (X^2 = 0.14, p = .939) based on our parametric bootstrap analysis (Figure 2A). Our mixed-effects model specification was nearly identical to that of recall, with the additions of question type as a fixed effect that distinguished detail and inference multiple-choice questions and a random intercept for multiple-choice question instead of idea unit. Consistent with our predictions for the individual difference measures, the model revealed a significant effect of reading ability (X^2 = 26.04, p < .001) and a significant effect of prior knowledge (X^2 = 10.94, p = .002).

There was some evidence that retrieval practice of main ideas impacted retention revealed by a significant interaction between participant group and prior knowledge ($X^2 = 7.05$, $p = .044$). A follow up simple slopes analysis revealed that the slope for RPm was not significantly different from zero ($B = -.02$, $p = .87$), but there were significant positive slopes for RPp ($B = .51$, $p < .01$) and NRP ($B = .44$, $p = .01$), indicating that low knowledge participants in the RPp and NRP groups performed worse than high knowledge participants. This suggests that main idea retrieval practice was beneficial for low knowledge participants.

Finally, there was evidence that good readers retained and comprehended the text more accurately than poor readers, indicated by a significant interaction between reading ability and question type ($X^2 = 5.22$, p = .024). Simple slopes analysis revealed a significant positive slope for reading ability in detail questions ($B = .29$, $p = .02$) and a stronger relationship in inference questions ($B = .64$, $p < .01$).

Experiment 2 Results

Control analyses: retrieval practice conditions boosted memory for practiced material

Experiment 2 analyses were identical to Experiment 1 and replicated our main findings that selective retrieval practice was not associated with RIFA or RIF. Once again, we ran analyses to rule out variables that could potentially confound between-group comparisons: the amount of time spent reading the text during part 1, individual differences in reading ability, and prior knowledge of the topic of the text. One way ANOVAs revealed no significant between group differences in reading time and prior knowledge (both Fs < 1, all p's > .7), however there was a significant main effect of participant group in reading ability ($F(2,187) = 3.43$, p = .035). Follow up pairwise comparisons did not detect significant differences (all t's < 2.3, all p's >

.06). These variables are included in all subsequent analyses to account for variance unrelated to the retrieval practice manipulation.

Once again, comparisons between the two retrieval practice groups revealed that, on average, those in the RPp group required more practice attempts than those in the RPm group in order to achieve criterion performance on the tested idea units (t (122) = 2.42, p = .017), but as we describe below, this both groups displayed robust testing effects.

Next, we examined retention of idea units that were practiced in either the RPm or RPp groups, which replicated the results of Experiment 1. participants in the RPm group showed better retention of the practiced main idea units than did participants in the RPp and NRP groups, who were not previously tested on these sections (Figure 1D). The parametric bootstrap analysis of main idea units revealed a significant effect of participant group (X^2 = 51.36, p < .001), and follow-up comparisons revealed significantly greater retention in the RPm group compared to the NRP group (z = 6.37, p < .001), but the same comparison for the RPp group was not significant (z = .53, p = .59). These results confirmed that selective retrieval of main idea units in the RPm group resulted in a testing effect.

Conversely, as shown in Figure 1E, participants in the RPp group showed better retention of the practiced peripheral idea units than did participants in the RPm and NRP groups, who were not previously tested on these sections. The bootstrap analysis for peripheral idea units revealed a significant effect of participant group (X^2 = 68.90, p < .001). Follow up comparisons revealed a significant effect of RPp compared to NRP (z = 6.18, p < .001), but the same comparison for the RPm group was not significant (z = .66, p = .51). These findings indicate that our testing manipulation enhanced retention of practiced material, which was consistent with our predictions and replicated Experiment 1.

Recall of non-practiced idea units: No overall evidence for RIFA or RIF, positive effects of reading ability and prior knowledge

We next turned to retention of idea units that were not previously tested in any of the three groups (Figure 1F). Contrary to our predictions, the bootstrap analysis of non-practiced idea units did not reveal a significant effect of participant group (X^2 = 0.67, p = .73). However, consistent with our predictions, we observed a significant effect of reading ability (B = .224, X^2 = 8.70, p = .008) and prior knowledge (B = .333, X^2 = 18.83, p < .001). Finally, we observed a significant three-way interaction between participant group, reading ability, and prior knowledge (X^2 = 8.66, p = .012). Figure 3 shows that among high knowledge participants, less-skilled readers showed RIF, and skilled readers showed RIFA, and among skilled readers, high knowledge participants showed RIFA, and low knowledge participants showed RIF. Low knowledge, less-skilled readers showed the worst recall performance.

We conducted Bayes Factor analyses in the same procedure as in Experiment 1 and for the same reasoning: to determine whether our data provided evidence for the null hypothesis, or were merely insensitive. The difference in retention between the RPm condition and the NRP condition yielded very strong evidence for the null hypothesis over the RIFA hypothesis (BF = .032) and insensitivity between the RIF and the null hypothesis (BF = 1.326). The difference in retention between the RPp condition and the NRP condition yielded strong evidence for the null hypothesis over the RIFA hypothesis (BF = .014) and insensitivity between the RIF and the null hypothesis (BF = .590). Despite prior studies showing RIF or RIFA effects, the bayes factor analyses here presented strong against the presence of a RIFA effect, replicating Experiment 1, and insensitivity to detect a RIF effect.

Multiple Choice: No significant main effect of participant group, positive effect of reading ability

Our predictions for the multiple-choice test were the same as in Experiment 1. We predicted that the RPm group would demonstrate a RIFA effect, and the RPp group would demonstrate a RIF effect. Furthermore, we predicted that the high knowledge participants and good readers would excel at inference multiple-choice questions. The bootstrap analysis failed to detect significant differences due to participant group (X^2 = 4.73 p = .122) or prior knowledge (X^2 = 4.02, p = .053); however, reading ability showed a strong positive relationship with multiple-choice performance (B = .619, X^2 = 51.8, p < .001).

Chapter 2 General Discussion

Here, we examined the effects of selective retrieval practice by having participants practice retrieval of main ideas or peripheral ideas before testing their retention for all of the text one day later. Based on prior findings (Chan, 2009), this paradigm–with high integration among passage concepts connected to main ideas–was well suited to find a RIFA effect in the main idea condition; conversely, the peripheral idea condition was predicted to show a RIF effect, based on the idea that retrieval competition causes RIF (Little et al., 2011). Our results showed that RIFA occurred only in Experiment 2, and only for high knowledge, skilled readers, but this result was not observed in Experiment 1. In contrast, there was some evidence for an overall RIF effect in Experiment 1's peripheral idea unit retrieval practice group, but Experiment 2 did not replicate this result. In Experiment 2 only, there was evidence of RIF in low-knowledge, skilled readers and in high knowledge, less-skilled readers. Overall, these results suggest that in the ecologically valid, cohesive science texts used here, selective retrieval practice produced inconsistent effects on non-practiced information that sometimes depended on reader characteristics. Below, we will

interpret our findings in relation to our predictions, the existing literature, and the situation model account, and we will conclude with the educational implications of our results.

We begin by reviewing the results of the recall test. Consistent with our predictions, in two experiments, we observed robust testing effects for practiced idea units in the main idea unit retrieval practice (RPm) and peripheral idea unit retrieval practice (RPp) conditions, when compared to the control group (NRP). Although these results are not surprising, they support our other results by showing that the critical selective retrieval practice manipulation was effective. Furthermore, we showed that learners with varying reading ability and prior knowledge all received the retention benefits of retrieval practice. To our knowledge, this is the first study to measure reading skill in an investigation of retrieval practice effects in educationally relevant materials, making this finding alone an important result.

Our primary focus concerned recall of non-practiced idea units, for which we predicted that the RPm group would demonstrate greater recall compared to the NRP group (RIFA) and the RPp group would demonstrate inferior recall compared to the NRP group (RIF). Contrary to these predictions, we did not observe any main effects of participant group in our analyses of non-practiced idea units. This was surprising because the literature has shown that cohesive text and one day retention intervals (Chan, 2009) and retrieval competition (Little et al., 2011) are factors that promote RIFA and RIF effects, respectively. Our experiments used texts that were rated to be more cohesive than those used by Chan (2009), yet our Bayes Factor analyses revealed strong evidence against a RIFA effect in the present experiments. We manipulated retrieval competition by creating one condition where participants practiced retrieval of main ideas (RPm), and one where peripheral ideas were practiced (RPp). We predicted that the RPp condition would demonstrate a RIF effect because peripheral ideas are unlikely to be well-

integrated into a coherent situation model of a text. The Bayes Factor analysis revealed moderate evidence in support of a RIF effect in Experiment 1, but this effect did not replicate in Experiment 2. These results suggest that the focal retrieval practice used here has very little effect beyond the practiced material.

We predicted that individual differences in reading ability and prior knowledge would moderate the effects of selective retrieval on non-practiced idea units. Both experiments revealed significant effects of reading ability, and Experiment 2 revealed a significant effect of prior knowledge. These results support our predictions and are consistent with the idea that participant variables that are likely to help in forming coherent situation models also predict retention of information one day later. Finally, Experiment 2 revealed a significant three-way interaction between participant group, reading ability, and prior knowledge, which shows that high knowledge, good readers demonstrated a RIFA effect, and both high knowledge, poor readers and low knowledge, good readers demonstrated RIF effects. This three-way interaction was not predicted and was not observed in Experiment 1, therefore it is weak evidence.

We predicted that good readers and those with greater prior knowledge would be more likely to show RIFA due to their greater ability to build coherent mental models. There was not consistent evidence that participants with different reading ability or prior knowledge experienced divergent effects of retrieval practice. The upshot of this absence of evidence is that less-skilled readers and those without prior knowledge did not experience deleterious effects of retrieval practice on retention of non-practiced information, which would be a major concern for the prescription of retrieval practice in educational contexts. In summary for the non-practiced idea units, we observed virtually zero evidence for RIFA effects, some evidence for a RIF effect, solid evidence for reading ability, and some evidence for prior knowledge.

Our predictions for the multiple-choice final test were the same as for the non-practiced idea units because the multiple-choice test questions were not specifically designed to probe the practiced idea units. Replicating the results of non-practiced recall, we did not observe significant differences in participant group in the multiple-choice test. We observed significant effects of reading ability in both experiments, replicating the results of the non-practiced idea units. Once again, we observed a significant effect of prior knowledge in Experiment 1, but it was only marginally significant in Experiment 2. Analyses of Experiment 1 revealed two interactions but they did not replicate in Experiment 2. In summary, the results of the multiple-choice final test largely replicated the results of the non-practiced idea units.

Despite the ubiquity of selective retrieval, only a few studies have investigated the consequences of selective retrieval for retention of educationally relevant text, and none have employed free recall to measure the full breadth of possible retention impacts; most used cued-recall criterial tests. We probed whether an ecologically valid form of competition produces RIFA or RIF in cohesive prose texts. Furthermore, given the critical role of reading ability and prior knowledge to forming a coherent mental model of a text, we predicted that these individual difference measures would moderate retention and explored whether they moderated any effects of selective retrieval. Previous studies have included only some of these characteristics, which hinders our ability to compare our results with these studies. Caroll et al. (2007) showed that expertise protects against RIF; however, their expert group were graduate students, which is likely a population of good readers, but reading ability was not measured. In Experiment 1, we did not replicate this result. In Experiment 2, there was some evidence for this finding such that there was a RIFA effect for high knowledge, skilled readers, and a RIF effect for high knowledge, less-skilled readers and for low knowledge, skilled readers. Chan (2009) observed a

RIFA effect after a one-day retention interval, in cohesive texts. As shown in Table 1, the text materials in the current study were considerably more cohesive than those used in previous studies, yet we failed to observe robust RIFA effects. A principal difference in the present study to Chan (2009) was that in that study, retrieval practice questions were constructed so as to facilitate retrieval of a complementary question. Finally, Little et al. (2011) argued that retrieval competition causes RIF, even in cohesive texts, by showing a RIF effect when distinct text materials were in competition. We manipulated competition, trying to create a naturalistic manipulation thereof using peripheral ideas, but there was minimal evidence that this produced RIF. Likewise, main idea retrieval practice had little effect beyond the testing effect for practiced units.

Why was retrieval practice so ineffective for non-practiced material in the current experiments? One possibility is that, by using cohesive texts and one-day delays, we biased our results toward RIFA, and without the additive benefit of facilitative questions, no RIFA effect was observed. Conversely, studies that observed RIF using educational texts did so by presenting the materials in a random order (Carroll et al., 2007; Chan, 2009), or by using a set of texts that were likely to interfere with one another (Little et al., 2011). We used a competition manipulation that we felt was typical of a testing situation on a single text but may have resulted in lower competition than in prior studies.

Chan (2009) put forward a situation model account of the effects of selective retrieval for non-practiced material, incorporating findings from a diverse set of stimuli. Their claim was that RIFA is most likely to occur when conditions encourage stimuli to be integrated into the same situation model, and RIF is most likely to occur when stimuli are spread out among many situation models. In the current study, we used text materials shown to score highly in cohesion

measures, yet we found very little evidence for RIFA. There are two noteworthy possibilities for why RIFA was not observed. First, if most participants could not integrate the texts into coherent situation models, then RIFA would not be expected. If the first scenario were true, then we would have observed poor performance on the final tests, particularly on the inference multiple-choice questions. In fact, final test performance on the multiple-choice test was very similar to the normative performance in the first study to use the current materials (Hinze et al., 2013). The second possibility is that the selective retrieval practice did not activate relevant situation models. In the studies reviewed here that demonstrated RIFA effects, there was always a one-to-one relationship between a tested and an non-practiced item. These items were essentially yoked together by experimental conditions that encouraged integration into the same situation model. In the current study, there was one practiced idea unit for every five to six non-practiced idea units, and there is no straightforward method to discern which idea units are integrated into which situation models. Perhaps RIFA would have occurred had participants practice tested more idea units. Future work could improve upon the current design and surmount these issues by including an initial open-ended retrieval practice phase, prior to selective retrieval practice. In this design, it would be clear for each participant which idea units were learned, so that they could be tested again after selective retrieval practice. Nevertheless, our approach used a highly naturalistic educational setting which offered improvements in ecological validity over prior research and was well suited to measure whether selective retrieval practice extends broadly to non-practiced material in these settings.

Finally, due to the fact that Experiments 1 and 2 did not always show the same results, further investigation into the influence of text characteristics on the sequelae of selective retrieval is merited. One difference between the two Experiments is that in Experiment 1, the text

was about viruses, during the COVID-19 pandemic, whereas the text in Experiment 2 covered the endocrine system. Despite the minor discrepancies in results between the two experiments, the relatively large sample size and rigor of our statistical analyses lends ample confidence in the validity of our results.

Retrieval practice has been widely recommended in educational contexts, however the sequelae of selective retrieval practice in prose texts are not well understood. Some studies have shown RIF effects, and others have shown RIFA effects. In the current study, we used cohesive text materials and an ecologically valid design, and we failed to find robust evidence for either RIFA or RIF. The implication of this result is that educators need not worry about deleterious effects of the type of focal retrieval practice used here, but they should not expect it to facilitate retention beyond the practiced material. We showed that across the range of reading ability and prior knowledge, learners experienced retention benefits for practiced material.

Chapter 2 Works Cited

Adesope, O. O., Trevisan, D. A., & Sundararajan, N. (2017). Rethinking the Use of Tests: A Meta-Analysis of Practice Testing. *Review of Educational Research*, *87*(3), 659–701. https://doi.org/10.3102/0034654316689306

Agarwal, P. K., Nunes, L. D., & Blunt, J. R. (2021). Retrieval Practice Consistently Benefits Student Learning: A Systematic Review of Applied Research in Schools and Classrooms. *Educational Psychology Review*, *33*(4), 1409–1453. https://doi.org/10.1007/s10648-021-09595-9

An Introduction to Mixed Models for Experimental Psychology. (2019). In *New Methods in Cognitive Psychology* (pp. 4–31). Routledge. https://doi.org/10.4324/9780429318405-2

Anderson, M. C. (2003). Rethinking interference theory: Executive control and the mechanisms of forgetting. *Journal of Memory and Language*, *49*(4), 415–445. https://doi.org/10.1016/j.jml.2003.08.006

Anderson, M. C., Bjork, R. A., & Bjork, E. L. (19950201). Remembering can cause forgetting: Retrieval dynamics in long-term memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *20*(5), 1063. https://doi.org/10.1037/0278-7393.20.5.1063

Anderson, M. C., & McCulloch, K. C. (1999). Integration as a general boundary condition on retrieval-induced forgetting. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *25*, 608–629. https://doi.org/10.1037/0278-7393.25.3.608

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, *67*, 1–48. https://doi.org/10.18637/jss.v067.i01

Carroll, M., Campbell-Ratcliffe, J., Murnane, H., & Perfect, T. (2007). Retrieval-induced forgetting in educational contexts: Monitoring, expertise, text integration, and test format. *European Journal of Cognitive Psychology*, *19*(4–5), 580–606. https://doi.org/10.1080/09541440701326071

Chan, J. C. K. (2009). When does retrieval induce forgetting and when does it induce facilitation? Implications for retrieval inhibition, testing effect, and text processing. *Journal of Memory and Language*, *61*(2), 153–170. https://doi.org/10.1016/j.jml.2009.04.004

Chan, J. C. K., McDermott, K. B., & Roediger III, H. L. (2006). Retrieval-induced facilitation: Initially nontested material can benefit from prior testing of related

material. *Journal of Experimental Psychology: General*, *135*, 553–571. https://doi.org/10.1037/0096-3445.135.4.553
Cohn-Sheehy, B. I., Delarazan, A. I., Crivelli-Decker, J. E., Reagh, Z. M., Mundada, N. S., Yonelinas, A. P., Zacks, J. M., & Ranganath, C. (2022). Narratives bridge the divide between distant events in episodic memory. *Memory & Cognition*, *50*(3), 478–494. https://doi.org/10.3758/s13421-021-01178-x
Dienes, Z. (2014). Using Bayes to get the most out of non-significant results. *Frontiers in Psychology*, *5*. https://www.frontiersin.org/articles/10.3389/fpsyg.2014.00781
van Dijk, T.A. & Kintsch, W. (1983). *Strategies of Discourse Comprehension*. Academic Press.
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
Gernsbacher, M. A., Varner, K. R., & Faust, M. E. (1990). Investigating differences in general comprehension skill. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *16*, 430–445. https://doi.org/10.1037/0278-7393.16.3.430
Graesser, A. C., McNamara, D. S., Louwerse, M. M., & Cai, Z. (2004). Coh-Metrix: Analysis of text on cohesion and language. *Behavior Research Methods, Instruments, & Computers*, *36*(2), 193–202. https://doi.org/10.3758/BF03195564
Graesser, A. C., Singer, M., & Trabasso, T. (1994). Constructing inferences during narrative text comprehension. *Psychological Review*, *101*, 371–395. https://doi.org/10.1037/0033-295X.101.3.371
Graesser, A., McNamara, D., & Louwerse, M. (2003). What readers need to learn in order to process coherence relations in narrative and expository text. In *Rethinking Reading Comprehension* (p. 82).
Halekoh, U., & Højsgaard, S. (2014). A Kenward-Roger Approximation and Parametric Bootstrap Methods for Tests in Linear Mixed Models – The R Package pbkrtest. *Journal of Statistical Software*, *59*, 1–32. https://doi.org/10.18637/jss.v059.i09
Hinze, S. R., Wiley, J., & Pellegrino, J. W. (2013). The importance of constructive comprehension processes in learning from tests. *Journal of Memory and Language*, *69*(2), 151–164. https://doi.org/10.1016/j.jml.2013.03.002
Jaeger, T. F. (2008). Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models. *Journal of Memory and Language*, *59*(4), 434–446. https://doi.org/10.1016/j.jml.2007.11.007
Jonker, T. R., Dimsdale-Zucker, H., Ritchey, M., Clarke, A., & Ranganath, C. (2018). Neural reactivation in parietal cortex enhances memory for episodically linked information. *Proceedings of the National Academy of Sciences*, *115*(43), 11084–11089. https://doi.org/10.1073/pnas.1800006115
Kintsch, W., & S, S. W. K., C. B. E. M. A. F. R. (1998). *Comprehension: A Paradigm for Cognition*. Cambridge University Press.
Little, J. L., Bjork, E. L., Bjork, R. A., & Angello, G. (2012). Multiple-Choice Tests Exonerated, at Least of Some Charges: Fostering Test-Induced Learning and Avoiding Test-Induced Forgetting. *Psychological Science*, *23*(11), 1337–1344. https://doi.org/10.1177/0956797612443370
Liu, X. L., & Ranganath, C. (2021). Resurrected memories: Sleep-dependent memory consolidation saves memories from competition induced by retrieval practice.

Psychonomic Bulletin & Review, *28*(6), 2035–2044. https://doi.org/10.3758/s13423-021-01953-6

MacGinitie, W. H., MacGinitie, R. K., Cooter, R. B., & Curry, S. (1989). Assessment: Gates-Macginitie Reading Tests, Third Edition. *The Reading Teacher*, *43*(3), 256–258.

MacLeod, M. D., & Macrae, C. N. (2001). Gone but Not Forgotten: The Transient Nature of Retrieval-Induced Forgetting. *Psychological Science*, *12*(2), 148–152. https://doi.org/10.1111/1467-9280.00325

McNamara, D. S. (2017). Self-Explanation and Reading Strategy Training (SERT) Improves Low-Knowledge Students' Science Course Performance. *Discourse Processes*, *54*(7), 479–492. https://doi.org/10.1080/0163853X.2015.1101328

McNamara, D. S., Kintsch, E., Songer, N. B., & Kintsch, W. (1996a). Are good texts always better? Interactions of text coherence, background knowledge, and levels of understanding in learning from text. *Cognition and Instruction*, *14*, 1–43. https://doi.org/10.1207/s1532690xci1401_1

McNamara, D. S., Kintsch, E., Songer, N. B., & Kintsch, W. (1996b). Are Good Texts Always Better? Interactions of Text Coherence, Background Knowledge, and Levels of Understanding in Learning From Text. *Cognition and Instruction*, *14*(1), 1–43. https://doi.org/10.1207/s1532690xci1401_1

McNamara, D. S., & Magliano, J. (2009). Chapter 9 Toward a Comprehensive Model of Comprehension. In *Psychology of Learning and Motivation* (Vol. 51, pp. 297–384). Academic Press. https://doi.org/10.1016/S0079-7421(09)51009-2

Murayama, K., Miyatsu, T., Buchli, D., & Storm, B. C. (2014). Forgetting as a consequence of retrieval: A meta-analytic review of retrieval-induced forgetting. *Psychological Bulletin*, *140*, 1383–1409. https://doi.org/10.1037/a0037505

O'Reilly, T., & McNamara, D. S. (2007). The Impact of Science Knowledge, Reading Skill, and Reading Strategy Knowledge on More Traditional "High-Stakes" Measures of High School Students' Science Achievement. *American Educational Research Journal*, *44*(1), 161–196. https://doi.org/10.3102/0002831206298171

O'reilly, T., & Mcnamara, D. S. (2007). Reversing the Reverse Cohesion Effect: Good Texts Can Be Better for Strategic, High-Knowledge Readers. *Discourse Processes*, *43*(2), 121–152. https://doi.org/10.1080/01638530709336895

Ozuru, Y., Dempsey, K., & McNamara, D. S. (2009). Prior knowledge, reading skill, and text cohesion in the comprehension of science texts. *Learning and Instruction*, *19*(3), 228–242. https://doi.org/10.1016/j.learninstruc.2008.04.003

Pinheiro, J. C., & Bates, D. M. (Eds.). (2000). Linear Mixed-Effects Models: Basic Concepts and Examples. In *Mixed-Effects Models in S and S-PLUS* (pp. 3–56). Springer. https://doi.org/10.1007/0-387-22747-4_1

R Core Team (2022). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Raaijmakers, J. G. W., & Jakab, E. (2013). Rethinking inhibition theory: On the problematic status of the inhibition theory for forgetting. *Journal of Memory and Language*, *68*(2), 98–122. https://doi.org/10.1016/j.jml.2012.10.002

Radvansky, G. A., & Zacks, R. T. (1991). Mental models and the fan effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *17*, 940–953. https://doi.org/10.1037/0278-7393.17.5.940

Roediger, H. L., & Karpicke, J. D. (2006). The Power of Testing Memory: Basic Research and Implications for Educational Practice. *Perspectives on Psychological Science*, *1*(3), 181–210. https://doi.org/10.1111/j.1745-6916.2006.00012.x

Rowland, C. A. (2014). The effect of testing versus restudy on retention: A meta-analytic review of the testing effect. *Psychological Bulletin*, *140*, 1432–1463. https://doi.org/10.1037/a0037559

Singmann, H., Bolker, B., Westfall, J., Aust, F., & Ben-Shachar, M. (2022). _afex: Analysis of Factorial Experiments_. R package version 1.2-0, <https://CRAN.R-project.org/package=afex>.

Singmann, H., & Kellen, D. (2019). An introduction to mixed models for experimental psychology. In New methods in cognitive psychology (pp. 4-31). Routledge.

Sweet, A. P., & Snow, C. E. (2003). *Rethinking Reading Comprehension*. Guilford Press.

Voss, J. F., & Silfies, L. N. (1996). Learning From History Text: The Interaction of Knowledge and Comprehension Skill with Text Structure. *Cognition and Instruction*, *14*(1), 45–68. https://doi.org/10.1207/s1532690xci1401_2

Wagenmakers, E.-J., Verhagen, J., Ly, A., Matzke, D., Steingroever, H., Rouder, J. N., & Morey, R. D. (2017). The Need for Bayesian Hypothesis Testing in Psychological Science. In *Psychological Science Under Scrutiny* (pp. 123–138). John Wiley & Sons, Ltd. https://doi.org/10.1002/9781119095910.ch8