The role of retrieval type and feedback in test-potentiated new learning

Matej Pavlić, Denis Vlašiček & Dragutin Ivanec

University of Zagreb, Faculty of Humanities and Social Sciences, Department of Psychology

Corresponding author:

Denis Vlašiček

Ivana Lučića 3, 10000 Zagreb, Croatia

dvlasice@ffzg.hr

Abstract

This study explored the effects of episodic and semantic retrieval, and feedback presentation, on learning of new complex material. Participants read a text divided into three parts, between which they engaged in (i) content-related (episodic retrieval) or (ii) general-knowledge testing (semantic retrieval), or (iii) reread the previous part. Participants in the two retrieval conditions were randomly assigned to either receive or not to receive feedback on their interpolated test achievement. Learning was measured through multiple choice questions whose distractors were designed specifically to enable capturing proactive interference. Planned analyses showed that participants in the episodic retrieval condition scored higher on the final test than participants in the other two groups. We found no evidence for an effect of feedback, nor for an interaction effect of interpolated activity type and feedback. Furthermore, no effect of either independent variable on the number of proactive intrusions was found, although the existence of an effect cannot be rejected, which is reflected in exploratory Bayesian analyses. Results are interpreted in terms of integration and metacognitive frameworks that have previously been suggested as explanations of the effect.

Keywords: test-potentiated new learning, interpolated activity, feedback, semantic retrieval, episodic retrieval

Introduction

The term "testing effect" refers to the finding that, when it comes to long-term retention of a piece of information, retrieving that information from memory trumps restudying it (Adesope, Trevisan, & Sundararajan, 2017; Glover, 1989; Karpicke & Roediger, 2008; Roediger III & Butler, 2011; Roediger III & Karpicke, 2006a, 2006b; Rowland, 2014). It is generally recognised that testing can have two types of effects — *direct* and *indirect* (Arnold & McDermott, 2013b; Roediger III & Karpicke, 2006a). Direct effects refer specifically to the increased retention that ensues from repeatedly *successfully* retrieving the target information — a process which is, presumably, reactivated at the time of a later test. A typical demonstration of the testing effect entails an initial learning phase, followed by a period during which participants either restudy the same material, engage in a memory test involving the studied material, or are not exposed to the original material at all. Finally, after a retention interval, an additional memory test reveals that the group subjected to a memory test during the intervening period has a distinct advantage over the other two groups.

On the other hand, indirect effects are brought about by some other process or processes besides the act of taking the test (Roediger III & Karpicke, 2006a; but for a different view, see Kornell, Klein, & Rawson, 2015). For example, *unsuccessful* retrieval attempts (which are not followed by feedback) can, through subsequent repeated encoding, also generate a testing effect, namely *test-potentiated (re)learning* (Arnold & McDermott, 2013a, 2013b; Izawa, 1966, 1970; Kornell, Hays, & Bjork, 2009; Wissman & Rawson, 2018). With the aim of disentagling test-potentiated (re)learning from the testing effect, Arnold and McDermott (2013b) let participants learn a list of 25 word pairs. One group completed nine cycles comprising a single test and a restudy session, while another completed three cycles comprising five tests and a restudy session. In order to isolate the effect of unsuccessful from successful retrieval attempts, they focused their analysis only on words not recalled on a test preceding a restudy episode. Results showed that, compared to taking fewer tests, taking more tests produces a greater increase in the proportion of *newly* retrieved items (i.e. words that were not retrieved on a pretest) in a test immediately following a restudy episode — a veritable potentiation of learning.

Test-potentiated new learning

Juxtaposed to the well established finding that attempting to recall studied material compared to restudying it, facilitates the long-term retention of *that* material, stand the results of a decade of research showing that retrieving previously studied information can even facilitate the acquisition of *new* information (Chan, Meissner, & Davis, 2018; Yang, Potts, & Shanks, 2018). If each additional study episode in the paradigm used to demonstrate test-potentiated learning contains *new* materials (giving a now standard blocked design; Chan, Manley, Davis, and Szpunar (2018)), one still observes that testing the memory of those new materials after each learning episode, compared to restudying the same materials, yields a greater number of correct responses and a decrease of proactive interference on a test administered to all subjects after the final learning episode (e.g. Szpunar, Khan, & Schacter, 2013; Szpunar, McDermott, & Roediger, 2008; Wissman, Rawson, & Pyc, 2011). Following the reasoning of Chan, Meissner, and Davis (2018), in this paper we will use the term "test-potentiated *new* learning" (TPNL) to denote this effect.

In one of the earliest studies showing the effect of TPNL, Darley and Murdock (1971) observed that, when recalling studied lists of words, participants systematically produce more prior-list intrusions when probed for a given list, if their memory of a prior list had not been tested before they proceeded to study the given list. These findings were corroborated by Tulving and Watkins (1974), who used an AB-AC interference paradigm, and found that omitting a test following the study of the AB list had a deleterious effect on learning AC items. Building on these results, Szpunar et al. (2008) conducted a study using a blocked design wherein they told their subjects to study five lists of items in anticipation of a final cumulative test. All subjects were tested immediately after studying the final list, but they engaged in different intermittent activities between studying the first four lists. One group was tested on each list after studying it, another group restudied each list, and a third group completed a mathematical distractor task. Participants whose memory was tested after each list produced more correct responses and fewer prior-list intrusions on the immediate test administered after studying the last list. The authors explained the found benefit of testing in terms of a segregation mechanism that prevents overburdening of retrieval cues, which, in the absence of testing, causes a build-up of proactive interference. The following decade has seen a renewal of interest in TPNL (Chan, Meissner, & Davis, 2018; Pastötter & Bäuml, 2014; Yang et al., 2018), with studies mainly using the multilist learning paradigm to delineate the scope of the effect with respect to various moderating variables: the relationship between materials to be studied in successive study episodes, varieties of study designs (blocked vs. interleaved), and populations, to name a few.

A particularly important question for our study and for real-world applications is whether these results generalise to materials more complex than word lists, and research conducted in the preceding decade mostly points to a positive answer (Divis & Benjamin, 2014; Jing, Szpunar, & Schacter, 2016; Szpunar, Khan, & Schacter, 2013; Wissman et al., 2011). Wissman et al. (2011) were the first to demonstrate TPNL with complex textual materials. In a series of five experiments, they presented their participants with prose passages, and showed that interpolating study episodes with attempts at free recall of the studied materials increased performance relative to a no-study condition and a condition where the interpolated activity involved solving mathematical equations. Interestingly, their results provided compelling evidence that release from proactive interference may not be the most appropriate explanation for TPNL, at least when it comes to complex learning materials. Extending these results to a more contemporary eduactional setting, Szpunar, Khan, and Schacter (2013) showed that TPNL occurs when the learning materials are online video lectures, and inclined towards a source monitoring account of the observed effect (Brewer, Marsh, Meeks, Clark-Foos, & Hicks, 2010). In the following section, we provide a short overview of heretofore proposed accounts of TPNL.

Theoretical overview

Recently, Chan, Meissner, and Davis (2018) provided a meta-analytic analysis and comprehensive overview of the literature, identifying four *nonconflicting* theoretical frameworks which were put forth throughout the years as viable explanations for TPNL. *Resource theories* generally posit that testing increases cognitive resources, but they propose different mechanisms by which this is achieved: (i) proactive interference reduction (e.g. Nunes & Weinstein, 2012; Szpunar et al., 2008; Wahlheim, 2015; Weinstein, McDermott, & Szpunar, 2011), (ii) restoration of encoding/attentional resources (e.g. Pastötter, Schicker, Niedernhuber, & Bäuml, 2011), or (iii) alteration of mind wandering patterns (e.g. Jing et al., 2016; Szpunar, Khan, & Schacter, 2013; Szpunar, Moulton, & Schacter, 2013). Whereas resource theories focus on the amount of deployable cognitive resources, *metacognitive theories* emphasise the optimisation of encoding strategies induced by retrieval attempts (e.g. Chan, Manley, et al., 2018; Cho, Neely, Crocco, & Vitrano, 2017). For example, in a recent investigation, Chan, Manley, et al. (2018) found that, compared to untested groups, the group whose memory for the first three word lists was subjected to interpolated testing displayed superior semantic organisation across lists. These findigs reflect a similar pattern obtained for the testing effect, where a greater number of tests is associated with improved organisation of output displayed upon testing (Karpicke, 2012; Zaromb & Roediger, 2010).

The key idea underlying the third framework — *context theories* — is that, apart from storing the studied information themselves, people store the related contextual information as well (e.g. Lehman, Smith, & Karpicke, 2014). Afterwards, the accessibility of this contextual information can affect the likelihood of successful retrieval of target information. Furthermore, the claim is that, unlike restudying, attempting retrieval causes an internal context change relative to the study context (Jang & Huber, 2008; Sahakyan & Kelley, 2002), and recalled items may be updated with contextual information from the retrieval attempt, while newly encountered information is still associated only with the study context. Therefore, recalling new-learning items is limited to only those items associated exclusively with the study context, providing them with the advantage observed upon testing. While this circumscription of separate learning episodes is at the core of both resource and context accounts, its effect on learning is supposedly different. According to the former, isolating a learning episode through attempts at recall increases resources for subsequent learning by preventing *encoding-based* proactive interference. On the other hand, the latter place the emphasis on later *retrieval* processes, whereby isolating an earlier learning episode reduces the memory search set for retrieval.

Finally, *integration theories* advance the notion that interpolated testing facilitates the integration of the new-learning material either with its retrieval cues or with the original-learning material. On one account, testing increases the likelihood of spontaneous covert retrieval of original-learning items during the study of new items, fostering their integration, thereby increasing conceptual organisation (e.g. Jing et al., 2016) and the effectiveness of retrieval cues (Pyc & Rawson, 2010). For example, Jing et al. (2016) found that interpolated testing increased the clustering of related information that is acquired across different segments within a video-recorded lecture.

Nonepisodic recall

One of the more curious findings in the field is that TPNL can arise not only after retrieving the previously studied material (episodic retrieval), but also after retrieval of information unrelated to the studied material from semantic memory (Divis & Benjamin, 2014; Pastötter et al., 2011), or from short-term memory (Pastötter et al., 2011), although there have been unsuccessful attempts at replication (e.g. Weinstein, McDermott, Szpunar, Bäuml, & Pastötter, 2015).

Pastötter et al. (2011) let their participants learn five lists of 20 words while engaging in varied interlist activities. They either restudied the lists, recalled the words from the list, generated as many words as they could from one of four semantic categories (e.g. professions), engaged in a 2-back short-term memory task, or counted backwards from a random three-digit number. They found that all three forms of retrieval induced TPNL. In their first experiment, Divis and Benjamin (2014) adapted the procedure from Pastötter et al. (2011), using only the semantic generation and distractor (counting backwards) tasks, and found that interleaved semantic retrieval enhanced performance for final list recall. They replicated and extended these findings in their second experiment by using complex learning materials: lists of words were replaced by texts related to animals, while learning was evaluated with short-answer and multiple-choice questions.

The argument these two groups of authors invoke to explain their results is that nonepisodic retrieval tasks sufficiently alter participant's internal context. Because the last study session is not affected by an additional context shift, a beneficial segregation of the final study context from the previous ones is produced, which reduces the memory search set. Chan, Meissner, and Davis (2018) conducted a metaregression, comparing the explanatory power of the four frameworks described above. Summarising their results, they highlighted resource and integration theories as accounts which have thus far garnered more empirical support, giving a slight upper hand to integration theories, while stating that context theories are least supported by extant research. Therefore, we opted to align our study design with the goal of comparing resource and integration frameworks.

Feedback

Although corrective feedback is known to augment the testing effect (Roediger III & Butler, 2011), there is a paucity of research into the effect of feedback on TPNL, especially when considering studies that have implemented the blocked design. Feedback is particularly important for recognition test such as multiple-choice tests since the usual benefit testing confers might turn into a disadvantage in case the test-taker selects a lure (Marsh, Roediger, Bjork, & Bjork, 2007; Roediger & Marsh, 2005). Moreover, evidence points to the timing of feedback being a relevant variable when gauging its influence on learning, with delayed feedback showing superior effects compared to immediate feedback (Butler, Karpicke, & Roediger, 2007; Butler & Roediger, 2008; Metcalfe, Kornell, & Finn, 2009; Smith & Kimball, 2010). For example, participants in a study by Butler and Roediger (2008) read prose passages and then either took or did not take an initial multiple-choice test. If they took the test, corrective feedback was either not given, given immediately after each answer was provided, or given in bulk after the entire test. A final test administered one week after the initial test revealed that, relative to studying, (1) taking an initial test alone tripled the success rate on the final test (22% performance increase), (2) giving immediate feedback on the initial test increased performance by 32%, but (3) that delayed feedback increased performance by 43%.

The variable of corrective feedback may be a fruitful avenue for research because resource and integration theories provide conflicting predictions regarding its effects on TPNL (Chan, Meissner, & Davis, 2018). Providing corrective feedback should increase the likelihood of intrusions during new learning, which are deemed beneficial from the standpoint of integration theories, but detrimental from the point of view of resource theories. Thus, feedback should reduce TPNL according to resource theories, but increase it according to integration theories.

Present study

Our study had two main goals. Firstly, we sought to replicate the TPNL effect in an ecologically valid setting, by using complex learning materials and standard multiple-choice items. Secondly, guided by the analysis of gaps in the field provided by Chan, Meissner, and Davis (2018), who identified a relative dearth of studies using nonepisodic retrieval and recognition (e.g. multiple-choice items) for the interpolated activity, and furthermore a lack of studies introducing feedback in a blocked study design, we attempted to expand the existing body of literature by employing a novel combination of variables, in order to examine their effects and interactions in the context of TPNL.

In particular, we assumed that retrieval could be the active component in interpolated activities that have been shown to give rise to TPNL. To test this, apart from using rereading as a control comparison task, we formed two memory tests, one of which tapped into episodic (assessing memory of the studied materials) while the other tapped into semantic (i.e. nonepisodic) memory (assessing general knowledge). Following the reasoning of Chan, Meissner, and Davis (2018), in order to pit integration and resources accounts of TPNL against each other, participants either were or were not given feedback upon completing an interpolated activity episode. Bearing in mind the necessity of systematically examining the impact of proactive interference on participants' performance, we used multiple-choice questions designed to assess memory both in terms of correct answers and susceptibility to intrusions.

Based on the preceding discussion, we predicted that, compared to the rereading group, participants in the retrieval groups would have a significantly higher average total score on the final test. Furthermore, we expected to find no difference between the two groups having different types of retrieval. With regards to proactive interference, we expected that participants engaging in episodic retrieval would display the lowest susceptibility to proactive interference, followed by participants in the semantic retrieval condition, and finally by the participants in the rereading condition. Looking at the two retrieval groups, we expected to find a significant main effect of feedback on the average number of correctly answered questions, as well as an interaction effect between feedback presentation and type of interpolated activity. Specifically, we assumed that presenting feedback would have a positive effect on the average number of correctly answered questions, but only for the participants in the content-related test condition. We also predicted that participants receiving feedback would have a significantly higher average number of intrusions than participants receiving no feedback. Finally, we expected to find an interaction effect of activity type and feedback presentation on the number of intrusions, but did not set a specific prediction regarding its pattern.

Methods

Participants and design

Undergraduate and graduate phonetics and psychology students (80.8% female, median age = 21, IQR = 3, range = [18, 31], total N = 207) participated in the study in exchange for course credit. We employed a 2 (interpolated activity: episodic vs semantic recall) x 2 (feedback: given vs not given) between-subjects design. Rereading served as a comparison interpolated activity, which was given to an additional control group. In total, this amounts to five separate groups, to which the participants were randomly assigned.

Materials and procedure

Participants read an expository text about weeds, drawn from a chapter in a university-level textbook. Some sentences and passages were slightly modified, so as to avoid odd language constructions; terms from the binomial nomenclature were translated, and, taking into account the characteristics of the target participant population, some plant names were removed from the text to make it more approachable. The text was divided into three interrelated parts (874, 754, and 835 words) constituting an integrated body of knowledge. Additionally, a practice text (768 words), not directly related to any of the other three parts, was taken from the same chapter. The materials were presented on a PC, in an application constructed using the open source *oTree* framework (version 2.1.35, Chen, Schonger, & Wickens, 2016) for the *Python* programming language (version 3.6.4, October 20, 2018).

For the interpolated activity, participants either (i) answered ten multiple choice questions related to the content of the part they have previously read (episodic recall, hereafter referred to as content-related testing), (ii) answered ten general knowledge multiple choice questions (semantic recall, hereafter referred to as general-knowledge testing) or (iii) reread the same part of the text they have previously read.

Further, we manipulated whether or not participants received feedback on their accomplishment on the interpolated tests. Feedback was presented on a separate screen which listed the questions, the participant's answers, and the correct answers in a tabular format. Incorrectly answered questions were highlighted in red, and correctly answered questions in green. After 40 seconds elapsed, a "Next" button appeared, allowing participants to proceed with reading the next part of the text. By setting this cooldown period, by emphasising that there would be a cumulative test, and by explicitely asking through written instructions, we wanted to encourage our participants to carefully examine the feedback. The feedback was presented for maximally 60 seconds, after which the application proceeded to the next part.

The general procedure is shown in Figure 1. Participants were first given a brief introduction to the study, and were encouraged to carefully read and follow the written instructions. Then, they were led to a computer which was running a fullscreen instance of the *oTree* application with a randomly chosen experimental condition. There, participants read the informed consent form and, in case there were no questions, started the experiment.

After entering their personal information, participants were presented with the instructions for their first task, which was to read the practice text at a speed that comes naturally to them. Unbeknownst to the participants, the time they took to read the practice text was recorded, and used as the basis for determining the reading time limits for the remaining texts. However, the lowest possible time limit was set to 5 minutes, and the longest to 8 minutes.

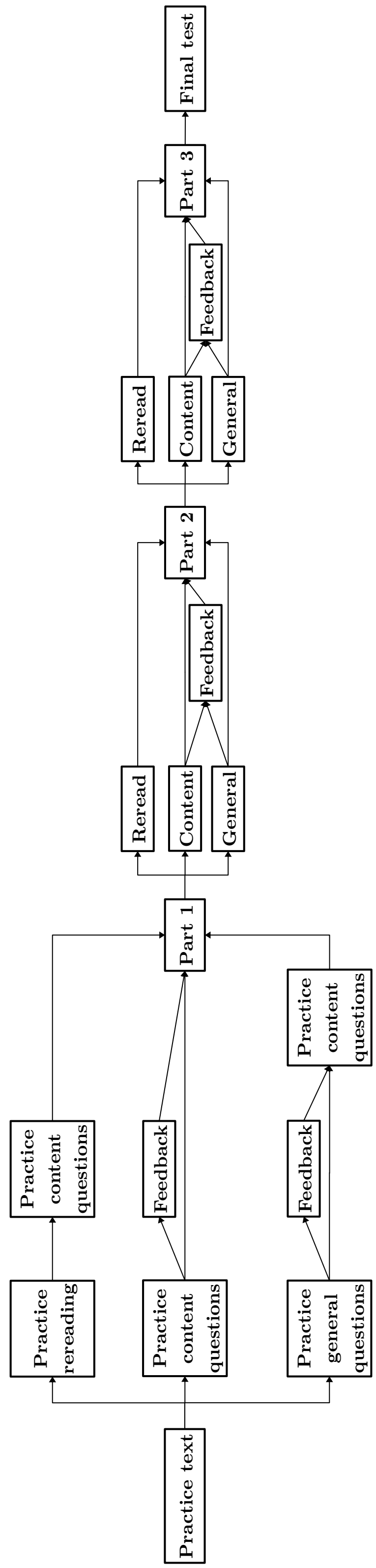

Figure 1. A flowchart depicting the experimental procedure.

Next, participants were familiarised with the interpolated activity they were going to perform during the main part of the procedure. The content-related test group answered four questions based on the practice text, the general-knowledge test group answered four general knowledge questions, and the rereading group reread the practice text (this time with the time limit applied). Subjects in the rereading and general knowledge conditions also answered the four questions related to the practice text, in order to familiarise themselves with the scope and specificity level of the questions they will receive after reading the final text. Participants assigned to the feedback condition also received feedback on their interpolated activity practice test achievement.

After the practice round, participants proceeded to the main part of the study, engaging in the interpolated activities they were assigned. Depending on the condition they were assigned to, they also received feedback after every interpolated test.

All participants were forewarned through initial instructions that there would be a cumulative test after the final part of the text, examining their knowledge of all three parts. In reality, the final test examined only the knowledge of the final part. Participants were presented with twenty questions examining their knowledge of that part. No feedback was presented after the final test, irrespective of the experimental condition. The computer recorded whether a participant correctly answered a question and whether the participant chose an intrusive distractor. This allowed us to compute our dependent variables — the total number of correct answers and the total number of intrusive distractors chosen.

In total, forty-four content related questions with four response options were generated from the presented parts of the text. Four questions were presented after the practice text, ten after each of the first two parts (only to the participants in the content related test condition), and twenty after the third part of the text (to all participants). Starting from the second ten-question-set, the distractor options were chosen so that (a) two distractors were plausible, but unrelated to the text, and (b) one distractor was a term or concept mentioned in the previous part of the text — this was considered to be the "intrusive" distractor (sometimes referred to as the "intrusor" in the rest of this article). Further, twenty-four general knowledge questions were generated. These questions were presented to participants

Table 1

Descriptive statistics for the DVs broken down by experimental condition.

Measure	Condition	n	M	SE_M	SD	min	max
Total correct	Content, feedback	41	13.22	0.508	3.25	2	19
	Content, no feedback	42	12.79	0.465	3.02	7	19
	General, feedback	40	10.97	0.533	3.37	1	17
	General, no feedback	40	10.47	0.449	2.84	5	16
	Rereading	40	10.88	0.443	2.80	4	17
Total intrusors	Content, feedback	41	3.15	0.258	1.65	0	7
	Content, no feedback	42	3.38	0.257	1.67	0	7
	General, feedback	40	4.17	0.318	2.01	0	8
	General, no feedback	40	4.58	0.288	1.82	1	9
	Rereading	40	4.62	0.350	2.21	1	10

in the general-knowledge test condition, after the first two parts of the text and after the practice text.

Results

Exclusion criteria

Prior to analysing the data, we excluded participants based on a priori set criteria. Participants who spent less than or equal to 90 seconds on the practice text were excluded (1 exclusion). Further, we wanted to exclude participants who had no correct answers on the final test (0 exclusions). Finally, we excluded participants who had stated that they had reading deficits (3 exclusions). This left us with a total sample of 203 participants. The descriptives for the sample are shown in Table 1.

Interpolated activity effect

Our first two hypotheses are concerned with the effects of different interpolated activities on the total number of correct answers and total number of intrusive distractors chosen. To test these hypotheses, we focused only on the groups which did not receive feedback (n = 122).

This was done because there was no feedback option for the rereading group, and we did not want to treat the feedback and no-feedback general-knowledge and content-related testing groups as equivalent without strong evidence supporting that assumption. We conducted a one-way MANOVA with interpolated activity as the independent variable and the total number of correct and intrusive options chosen as dependent variables. The correlation between our DVs calculated on the whole sample is $r(201)$ = -.707 (95% CI: [-.77, -.63], $p <$ .0001).

Pillai's V for the analysis is .126, p = .004 (Wilks' Λ = .875, p = .003). The effect size, calculated as ω^2_{mult} = .109 (bootstrap median[1] = .132, BC_α 95% CI = [.012, .201]). To further inspect the relationship of the interpolated activities with our dependent variables, we conducted a Roy-Bargmann stepdown analysis, as suggested by Tabachnick and Fidell (2012; a linear discriminant analysis with the same aim is available in the supplementary materials). The total number of correct answers was a priori chosen to be the higher priority variable. According to Tabachnick and Fidell (2012), the higher priority variable can be chosen based on theoretical or practical grounds. Since the total number of correct answers is the criterion that determines a student's success in a testing context, we chose this dependent variable as the higher priority one. Therefore, we first conducted an ANOVA with interpolated activity type as the independent variable and the total number of correct answers as the dependent variable.

As could be expected, the ANOVA points to an interpolated activity effect, with $F(2, 119)$ = 7.541, p = .001. Following the ANOVA, we conducted an ANCOVA, with the total number of correct answers as the covariate, and the total number of intrusors as the dependent variable. The results imply a main effect of the total number of correct answers ($F(1, 118)$ = 79.674, $p <$ .0001), but after we took into account the number of correct answers, we found no evidence for an effect of interpolated activity on the total number of chosen intrusors ($F(2, 118)$ = 0.844, p = .433). Thus far, results point to a lack of evidence to support our second hypothesis that the type of interpolated activity will have an effect on the number of intrusors.

[1] All bootstrap estimates taken from 10000 replications.

In order to test our first hypothesis, we contrasted (i) the rereading group with the two test groups, and (ii) the two test groups with each other, taking only the total number of correct answers as the DV. The first contrast found no evidence of a difference between the rereading group and the two test groups ($t(119) = 1.355$, $p = .178$, $g_s = 0.19$, 95% CI = [-0.19, 0.57], Cohen's $U_{3,g_s} = 57.6\%$, probability of superiority = 55.39%). However, there was a difference between the two test groups ($t(119) = 3.62$, $p = .0004$, $g_s = 0.66$, 95% CI = [0.21, 1.1], Cohen's $U_{3,g_s} = 74.43\%$, probability of superiority = 67.88%). Participants in the content related test group scored higher on the final test than participants in the general knowledge test condition. These two findings are not in line with our predictions.

The interaction between feedback and interpolated activity type

The remaining hypotheses deal with the effect of feedback on the total number of correct answers and the total number of intrusors. Therefore, these analyses were carried out on the data from participants in the general and content related test conditions only ($n = 163$). To test these hypotheses, we first conducted a two-way MANOVA with interpolated activity and feedback as independent variables, and total number of correct answers and total number of intrusors as the dependent variables.

Pillai's V for the interpolated activity effect (calculated with type III sums of squares) is .071, $p = .003$ (Wilks' $\Lambda = .929$, $p = .003$) confirming the main effect of interpolated activity type. The effect size $\omega^2_{mult} = .065$ (bootstrap median = .072, BC_α 95% CI = [.007, .139]).

On the other hand, we found no evidence for an effect of giving feedback on the linear combination of our two dependent variables — Pillai's V = .003, $p = .800$ (Wilks' $\Lambda = .997$, $p = .800$). The effect size is $\omega^2_{mult} = -.003$ (bootstrap median = .003[2]).

Furthermore, we found no evidence for an interaction effect between activity type and feedback — Pillai's V = .001, $p = .941$ (Wilks' $\Lambda = .999$, $p = .941$). The effect size $\omega^2_{mult} = -.005$ (bootstrap median = .003[3]). Both the feedback and the interaction estimates of ω^2_{mult}

[2] The BC_α 95% CI for this estimate is $[-.006, .004]$.

[3] The BC_α 95% CI = $[-.006, -.005]$. Our guess is that this odd result is due to the fact that most of the density is concentrated around 0, causing an unreliable estimate. The same could be said for the CI in footnote 2.

Table 2

ANOVA and ANCOVA models for the second Roy-Bargmann procedure.

Term	*SS*	*df*	*F*	*p*
ANOVA				
Activity	109.393	1	11.200	.001
Feedback	3.904	1	0.400	.528
Activity x Feedback	0.045	1	0.005	.946
Residuals	1553.046	159		
ANCOVA				
Activity	0.301	1	0.175	.676
Feedback	0.173	1	0.100	.752
Total correct	63.216	1	36.760	$<$ .0001
Activity x Feedback	0.813	1	0.473	.493
Activity x Total correct	0.862	1	0.501	.480
Feedback x Total correct	0.130	1	0.075	.784
Activity x Feedback x Total correct	1.229	1	0.715	.399
Residuals	266.551	155		

are to be considered to be zero, given their negative values.

Again, we conducted a follow-up Roy-Bargmann stepdown analysis. In the ANOVA model with the total number of correct answers as the dependent variable and the type of interpolated activity, feedback and their interaction as predictors, only the type of activity seems to be relevant ($F(1, 159) = 11.2, p = .001$). This result also shows that participants in the content related test condition scored higher on the final test than the participants in the general knowledge test condition, which should be no surprise given the results of the first stepdown analysis. In the second step, we fit an ANCOVA model with the total number of correct answers as the covariate. In this model, the type of interpolated activity ceases to be a relevant predictor ($F(1, 155) = 0.175, p = .676$). The full models are shown in Table 2.

To summarise, contrary to our expectations, we find no evidence of an effect of feedback on the total number of correctly answered questions. Also, we found no evidence for an interaction effect of feedback and type of interpolated activity on the total number of correct answers. The same findings apply to the predictions regarding the total number of intrusors chosen.

Additional analyses

Because it is theoretically interesting to see whether there is evidence for absence of a difference between certain conditions, or no effect of certain manipulations, we conducted a Bayesian reanalysis of the two Roy-Bargmann stepdown procedures. Since these analyses had not been planned, we decided to use the default priors provided in the *BayesFactor* (Morey & Rouder, 2018) package.[4]

Bayesian reanalysis of the first Roy-Bargmann procedure

As was earlier done in a frequentist setting, we first fit an ANOVA model with the total number of correct answers as the dependent variable, and the type of interpolated activity as the predictor. All effects are expressed as deviations from the estimated posterior subsample mean of 11.381. The estimated mean of the effect of content related testing is 1.254 (95% HDI = [0.553, 2.005]). The 95% highest density interval of the posterior indicates that there is a fair amount of uncertainty around the exact magnitude of the effect of content-related testing. However, most of the probability density is quite far above zero, implying that there really is a positive effect. The means of the posterior distributions for the general-knowledge-test and rereading conditions *b*s are -0.805 (95% HDI = [-1.549, -0.116]) and -0.449, (95% HDI = [-1.125, 0.257]) respectively. Most of the posterior distribution for the effect of general knowledge testing lies below zero, pointing to a negative effect on the total number of correct answers, although the distance is not as marked as in the content-related condition. On the other hand, there is a lot of uncertainty about the effect of rereading, compared to the other two estimates. Still, 89.8% of the posterior lies below zero, which lead us to believe that the effect is most likely negative.

Furthermore, we wanted to explore the difference between the rereading and general-knowledge-test conditions, given their somewhat similar coefficient and HDI estimates, as well as sample means. To do this, we conducted a Bayesian t-test, again with the *BayesFactor* package's default priors. The estimated posterior mean of the difference in the total number of correct answers between the two groups is -0.362 (95% HDI = [-1.49, 0.856]). As can be

[4] All posteriors obtained from 6000 simulations.

seen from the HDI, there is a lot of uncertainty around the estimate of the difference, which points to a lack of evidence for any claim regarding the effect.

In the second step of the Roy-Bargmann procedure, we fit an ANCOVA model with the total number of correct answers as the covariate and the total number of intrusive options chosen as the dependent variable. Effects are again expressed relative to the estimated posterior subsample mean of 4.193. There is uncertainty around the estimates of the effects of the different experimental conditions — content related testing b = -0.214 (95% HDI = [-0.583, 0.146]), general-knowledge testing b = 0.072 (95% HDI = [-0.288, 0.424]), rereading b = 0.142 (95% HDI = [-0.216, 0.494]). The HDIs show that there could be either a slight increase or a slight decrease in the number of intrusors, which prevented us from making a conclusion about the nature of the effects. However, given the current data and priors, we find the following — 87.43% of the posterior for the effect of content related testing falls below zero; 65.57% of the posterior for the effect of general knowledge testing falls above zero; 77.68% of the posterior for the effect of rereading falls above zero. Given the stated, there is some evidence implying that content related testing decreases the number of intrusors chosen, after controlling for the effect of the total number of correct answers. Further, there is some, albeit weaker evidence that rereading leads to an increase in the number of chosen intrusive distractors. Lastly, the posterior of the general knowledge testing effect points to no particular direction. A stronger test of these claims is desired.

Bayesian reanalysis of the second Roy-Bargmann procedure

In the second Roy-Bargmann analysis, we wanted to test whether there is an effect of the type of interpolated activity, receiving feedback, and their interaction on the total number of correct answers and chosen intrusors. Again, we first fit an ANOVA model with the two predictors and the total number of correct answers as the dependent variable.

Effects are expressed relative to the estimated posterior subsample mean of 11.868. We found that content related testing leads to an increase in the total number of correct answers, b = 1.086 (95% HDI = [0.589, 1.559]), compared to the general knowledge testing. This is aligned with the finding obtained in the frequentist setting. The mean of the posterior for the effect of receiving feedback is 0.218 (95% HDI = [-0.251, 0.679]). The HDI around the estimate precludes any firm conclusions regarding the effect of receiving feedback. However, we will mention that 82.25% of the posterior lies above zero, implying a possible positive effect on learning. Finally, the estimate for the interaction effect (being in the content condition and receiving feedback) is -0.013 (95% HDI = [-0.46, 0.432]). This could point to there not being a relevant interaction effect. According to the collected data and the priors, we could claim that the effect is practically equivalent to zero if we were not interested in a half-point increase or decrease in the average scores (i.e. defining a region of practical equivalence (ROPE) between [-0.5, 0.5]). Still, greater precision, which would require further data collection, is desired.

We continue with the ANCOVA model, taking the total number of correct answers as the covariate. The estimate of the intercept is 3.821 (95% HDI = [3.6, 4.03]). The estimate for the effect of content related testing on the total number of intrusive distractors chosen is b = -0.118 (95% HDI = [-0.325, 0.092]), compared to general knowledge testing. There is some evidence for a slight decrease in the number of intrusive distractors chosen in the content related testing condition. However, an increase is also possible, but less likely and negligibly small. The estimate for the effect of receiving feedback is -0.091 (95% HDI = [-0.302, 0.121]). Although the mean of the posterior is close to zero, the lower bound of the HDI shows that values which may be considered non-negligible are still somewhat probable. Therefore, we shall refrain from making a judgement regarding the effect of feedback on choosing intrusive distractors. Finally, the estimate of the interaction effect is b = 0.047 (95% HDI = [-0.153, 0.244]). The mean of the posterior is close to zero, and we could declare the effect to be practically equivalent to zero with a ROPE of approximately [-0.25, 0.25].

As previously stated, all these analyses were not planned a priori. This warrants certain caveats. The *BayesFactor* package's default priors were used. The appropriateness of these priors should certainly be questioned. However, we decided to use them because we did not want to choose priors after already seeing the data, which would have been more problematic. Further, the statements about effects made in this section are noncommittal. Whether a 0.5 increase or decrease in the total number of correct answers is practically equivalent to zero or not is left to the reader.

Deviations from the preregistered analysis plan

Initially, we had planned to do a robustness check of our findings using data with an additional exclusion criterion, based on the number of times each participant had read each of the three parts of the main text. This analysis was never conducted because (i) applying this criterion would have lead to unacceptably low power and (ii) the participants' estimates of the number of times they had read each part were similarly distributed across all conditions. Further, we had planned to conduct a TOST procedure to test whether there is no difference between the content-related and general-knowledge testing groups. This analysis was not conducted because we did find a difference. A Bayesian t-test was also considered for the same comparison, but was dropped early on due to some conceptual concerns.

Discussion

The aim of this study was to explore the effects of different interpolated activities and feedback reception on learning complex materials. Participants read three prose passages, and engaged in interpolated activities between reading episodes. Two testing activities were chosen so as to tap into episodic (content related testing) or semantic (general knowledge testing) memory, while a rereading condition served as a control. Participants in the episodic and semantic retrieval conditions were also randomly assigned to receive or not to receive feedback. Learning was measured through the total number of correct answers and the number of intrusive distractors chosen on a test assessing memory of the final part.

We found evidence for an effect of interpolated activity type on TPNL — treating the two dependent variables as manifestations of TPNL, we conducted a MANOVA, revealing that participants engaging in episodic retrieval exhibited greater TPNL than both participants who engaged in semantic retrieval and those in the control condition. Moreover, a Roy-Bargmann stepdown analysis was used to tease apart the contributions of the two dependent variables to the found effect, and it showed that observed differences were driven primarily by the number of correct responses, while we found no evidence for the contribution of proactive interference. These results are not entirely in line with extant research. While our results point to an exclusive role of episodic retrieval in TPNL, data from Pastötter et al. (2011) suggest that retrieval from both long-term and short-term memory can generate the effect. Notably, Pastötter et al. (2011) did find that the effect of various types of retrieval depends on the performance measure under examination. In their study, when considering the number of correct recalls, three different types of recall produced a comparable level of TPNL. However, while both episodic and semantic retrieval effectively eliminated intrusions during final recall, short-term memory retrieval led to intrusion rates that were equivalent to those observed in the no-retrieval conditions, suggesting a possible substantive difference between types of retrieval. It is important to point out that the effects observed by these authors were obtained by using learning materials simpler than the ones we used, which was shown to be an important moderating variable (Chan, Meissner, & Davis, 2018).

The few studies that have suggested that nonepisodic forms of recall may serve as effective methods of learning potentiation have drawn on context and resource theories to explain their results (Divis & Benjamin, 2014; Pastötter et al., 2011). Divis and Benjamin (2014) proposed that retrieval processes enhance context fluctuation, associating specific context cues with specific study episodes, thereby increasing the contextual disparity between information acquired across study sessions. This, in turn, isolates individual learning episodes, and reduces the memory search set and proactive interference. The lack of evidence for an effect of semantic retrieval on learning in our study may be taken as an argument against this encoding "resetting" and context change accounts of TPNL because, presumably, semantic retrieval should have produced the internal context change required for resetting the encoding process (Pastötter et al., 2011). Finally, we will mention that the Bayesian estimate of the effect of nonepisodic recall lends support for the claim that it does not enhance learning.

While we found no evidence for an effect of feedback on TPNL, the exploratory Bayesian analysis does not exclude the possibility of a feedback effect. Still, the estimates we have obtained point to an effect which could be practically equivalent to zero, i.e. insignificantly small for real-world purposes. From this we gather that the collected data provide no evidence that a proactive interference reduction mechanism underpins TPNL. Interpreting these results warrants caution, though, since a more precise estimate of the effect is desirable.

Importantly, our choice of learning materials could have prevented us from finding evidence in favour of context theories and an account based on the reduction of proactive interference. To our knowledge, only one study (Divis & Benjamin, 2014, Experiment 2) investigating the impact of different types of retrieval on TPNL used complex learning materials, but it did not allow for a direct assessment of the level of proactive interference. Previous work has shown that release from proactive interference may play basically no role when it comes to learning complex materials. Aiming primarily to replicate the basic pattern of findings of Szpunar et al. (2008), Wissman et al. (2011) used prose passages as learning materials, and found a very low baseline level of proactive interference in the control group, which effectively excluded an explanation based on release from proactive interference. In line with results obtained by Wissman et al. (2011), who used free recall as the testing format, the recognition-level method which we employed in order to examine possible interference effects showed no signs of proactive interference beyond those one could have expected to occur by chance alone.

Although our results do permit setting aside some accounts within the resource theoretical and context change frameworks, other theories still provide viable explanations of TPNL. Wissman et al. (2011) fashioned an account very much in line with the general ideas behind integration theories. Specifically, they proposed that interpolated testing induces a stronger activation and retention of learned information, whose accessibility further facilitates comprehension and encoding of new related materials, especially connected discourse and lecture videos. More recent studies provided supportive evidence for explanations relying on changes in patterns of mind wandering (Szpunar, Khan, & Schacter, 2013), whereby testing increases mind wandering related to the acquired information (Jing et al., 2016). Wissman et al. (2011) suggested an additional nonconflicting metacognitive explanation based on encoding strategy changes, mediated by possible failures of retrieval (Bahrick & Hall, 2005), whereby subjects use immediate feedback on recall to adjust their encoding strategy. In line with these proposals, recent studies have shown that performing retrieval modifies the learner's approach to new information (Cho et al., 2017; Soderstrom & Bjork, 2014), which may lead to superior semantic organisation of acquired knowledge (Chan, Manley, et al., 2018; Jing et al., 2016). However, our opinion is aligned with that of Chan, Manley, et al.

(2018), who caution that the application of a strategy change account to TPNL of complex learning materials should be preceded by an adequate description and operationalisation of an "advantageous" encoding strategy, which would enable precise tests of the proposed mechanism.

Finally, we have to address certain methodological concerns. In our study participants were thoroughly informed regarding the activities they would encounter during the procedure, including the final test that followed the last reading episode. The typical instruction given to participants in the TPNL paradigm is that interpolated activities will be determined randomly by the computer Yang et al. (2018). Thus, an attempt is made to equalise the expectations of a final test across conditions, and to ensure continued processing of materials across the study sequences. Nevertheless, learners dynamically adjust their expectations based on their experiences of the experimental procedure, regardless of being told that the activities they are given are determined randomly (Weinstein, Gilmore, Szpunar, & McDermott, 2014). If they take a test, they will more likely expect another one, and such expectations have been shown to make a significant difference for how one approaches the encoding task (e.g. Szpunar, McDermott, & Roediger, 2007). Moreover, a basic assumption we have made is that the interpolated activity that served the function of activating retrieval from semantic memory was effective.

Further, we cannot exclude the possibility that the nature of our interpolated activities had differential effects on our participants' motivation. Several participants did remark that the text was tedious, and it is possible that the motivation of participants in the episodic retrieval condition persisted throughout the procedure, while that of the other participants waned as the procedure progressed. However, if this were true, we can easily imagine that participants in the rereading condition should have obtained the lowest scores, given that they were instructed to read the texts twice. However, this is not the case. Furthemore, we believe that most people would agree that answering general knowledge questions is more interesting than answering questions about weeds, and that this activity would, therefore, help sustain the motivation level. On the other hand, differences in motivation could have been caused by unequal task difficulties — the mean proportions of correctly answered questions are larger in the content-related than in the general-knowledge testing conditions. Importantly, the mean proportion of correct answers on the first interpolated content-related test is higher than on the second. If the tests were equally difficult, and if there were a TPNL effect, we would expect higher scores on the second test. This points to the tentative conclusion that the interpolated tests themselves differ in difficulty. Thus, we cannot reject the possibility that differing difficulties have had an effect on our participants' achievement. However, Divis and Benjamin (2014) argue that the difficulty of the interpolated tasks is irrelevant. Still, such claims are yet to be corroborated by experimental data.

To conclude, our findings confirm the effect of episodic recall on TPNL, but we fail to find evidence for an effect of semantic recall. Further, evidence for an effect of feedback is also lacking. Our data are generally aligned with predictions stemming from metacognitive and integration theories of TPNL, and speak against proactive interference reduction accounts within the wider framework of resource theories.

Notes

Analyses conducted using the *R* language (R Core Team, 2019). Plots created using *ggplot2* (Wickham, 2016). Bootstrap conducted using the *boot* package (Canty & Ripley, 2017). Methods and analyses written using *rmarkdown* (Allaire et al., 2019) and *knitr* (Xie, 2019). The package *car* (Fox & Weisberg, 2011) was used to obtain type III sums of squares. *compute.es* (Re, 2013) was used to obtain effect sizes for contrasts. *kableExtra* was used to help generate tables (Zhu, 2019). Other utilities used are *tidyverse* (Wickham, 2017), *magrittr* (Bache & Wickham, 2014), *here* (Müller, 2017), *conflicted* (Wickham, 2018), *psych* (Revelle, 2018). Highest density intervals obtained using *HDInterval* (Meredith & Kruschke, 2018).

Funding information

This study was conducted under the project *E-rudito: An advanced online educational system for smart specialization and jobs of the future* (KK.01.2.1.01.0009.), which is funded from the European Regional Development Fund, Operational programme competitiveness and cohesion 2014–2020.

Open Practices Statement

The analysis plan was preregistered on GitHub (`analysis-plan.md`; first commit of analysis plan: `b101f42`; final relevant commit: `16afea3`), as were the hypotheses (`design.md`; first commit of hypotheses: `b101f42`; final commit: `dd0f863`). The repository also serves all project materials, data and analyses scripts, together with the whole project history. It can be found at `github`. Materials are also available at `osf`.

References

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. doi: 10.3102/0034654316689306

Allaire, J. J., Xie, Y., McPherson, J., Luraschi, J., Ushey, K., Atkins, A., ... Iannone, R. (2019). Rmarkdown: Dynamic Documents for R. [Computer software]. Retrieved from `https://CRAN.R-project.org/package=rmarkdown` (Version 1.12)

Arnold, K. M., & McDermott, K. B. (2013a). Free recall enhances subsequent learning. *Psychonomic Bulletin & Review*, *20*(3), 507-513. doi: 10.3758/s13423-012-0370-3

Arnold, K. M., & McDermott, K. B. (2013b). Test-potentiated learning: Distinguishing between direct and indirect effects of tests. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *39*(3), 940-945. doi: 10.1037/a0029199

Bache, S. M., & Wickham, H. (2014). magrittr: A Forward-Pipe Operator for R [Computer software]. Retrieved from `https://CRAN.R-project.org/package=magrittr` (Version

1.5)

Bahrick, H. P., & Hall, L. K. (2005). The importance of retrieval failures to long-term retention: A metacognitive explanation of the spacing effect. *Journal of Memory and Language*, *52*(4), 566-577. doi: 10.1016/j.jml.2005.01.012

Brewer, G. A., Marsh, R. L., Meeks, J. T., Clark-Foos, A., & Hicks, J. L. (2010). The effects of free recall testing on subsequent source memory. *Memory*, *18*(4), 385-393. doi: 10.1080/09658211003702163

Butler, A. C., Karpicke, J. D., & Roediger, H. L. (2007). The effect of type and timing of feedback on learning from multiple-choice tests. *Journal of Experimental Psychology: Applied*, *13*(4), 273-281. doi: 10.1037/1076-898X.13.4.273

Butler, A. C., & Roediger, H. L. (2008). Feedback enhances the positive effects and reduces the negative effects of multiple-choice testing. *Memory & Cognition*, *36*(3), 604-616. doi: 10.3758/MC.36.3.604

Canty, A., & Ripley, B. D. (2017). boot: Bootstrap R (S-Plus) Functions [Computer software]. Retrieved from `https://CRAN.R-project.org/package=boot` (Version 1.3-20)

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

Chan, J. C. K., Meissner, C. A., & Davis, S. D. (2018). Retrieval potentiates new learning: A theoretical and meta-analytic review. *Psychological Bulletin*, *144*(11), 1111-1146. doi: 10.1037/bul0000166

Chen, D. L., Schonger, M., & Wickens, C. (2016). oTree—An open-source platform for laboratory, online, and field experiments. *Journal of Behavioral and Experimental Finance*, *9*, 88-97. doi: 10.1016/j.jbef.2015.12.001

Cho, K. W., Neely, J. H., Crocco, S., & Vitrano, D. (2017). Testing enhances both encoding and retrieval for both tested and untested items. *Quarterly Journal of Experimental Psychology*, *70*(7), 1211-1235. doi: 10.1080/17470218.2016.1175485

Darley, C. F., & Murdock, B. B. (1971). Effects of prior free recall testing on final recall and recognition. *Journal of Experimental Psychology*, *91*(1), 66-73. doi: 10.1037/h0031836

Divis, K. M., & Benjamin, A. S. (2014). Retrieval speeds context fluctuation: Why semantic generation enhances later learning but hinders prior learning. *Memory & Cognition*, *42*(7), 1049-1062. doi: 10.3758/s13421-014-0425-y

Fox, J., & Weisberg, S. (2011). *An R Companion to Applied Regression* (Second ed.). Thousand Oaks CA: Sage.

Glover, J. A. (1989). The "Testing" Phenomenon: Not Gone but Nearly Forgotten. *Journal of Educational Psychology*, *81*(3), 392-399.

Izawa, C. (1966). Reinforcement-Test Sequences in Paired-Associate Learning. *Psychological Reports*, *18*(3), 879-919. doi: 10.2466/pr0.1966.18.3.879

Izawa, C. (1970). Optimal potentiating effects and forgetting-prevention effects of tests in paired-associate learning. *Journal of Experimental Psychology*, *83*(2, Pt.1), 340-344. doi: 10.1037/h0028541

Jang, Y., & Huber, D. E. (2008). Context retrieval and context change in free recall: Recalling from long-term memory drives list isolation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *34*(1), 112-127. doi: 10.1037/0278-7393.34.1.112

Jing, H. G., Szpunar, K. K., & Schacter, D. L. (2016). Interpolated testing influences focused attention and improves integration of information during a video-recorded lecture. *Journal of Experimental Psychology: Applied*, *22*(3), 305-318. doi: 10.1037/xap0000087

Karpicke, J. D. (2012). Retrieval-Based Learning: Active Retrieval Promotes Meaningful Learning. *Current Directions in Psychological Science*, *21*(3), 157-163. doi: 10.1177/0963721412443552

Karpicke, J. D., & Roediger, H. L. (2008). The Critical Importance of Retrieval for Learning. *Science*, *319*(5865), 966-968. doi: 10.1126/science.1152408

Kornell, N., Hays, M. J., & Bjork, R. A. (2009). Unsuccessful retrieval attempts enhance subsequent learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *35*(4), 989-998. doi: 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*(1), 283-294. doi: 10.1037/a0037850

Lehman, M., Smith, M. A., & Karpicke, J. D. (2014). Toward an episodic context account of retrieval-based learning: Dissociating retrieval practice and elaboration. *Journal of*

Experimental Psychology: Learning, Memory, and Cognition, *40*(6), 1787-1794. doi: 10.1037/xlm0000012

Marsh, E. J., Roediger, H. L., Bjork, R. A., & Bjork, E. L. (2007). The memorial consequences of multiple-choice testing. *Psychonomic Bulletin & Review*, *14*(2), 194-199. doi: 10.3758/BF03194051

Meredith, M., & Kruschke, J. (2018). HDInterval: Highest (Posterior) Density Intervals [Computer software]. Retrieved from `https://CRAN.R-project.org/package=HDInterval` (Version 0.2.0)

Metcalfe, J., Kornell, N., & Finn, B. (2009). Delayed versus immediate feedback in children's and adults' vocabulary learning. *Memory & Cognition*, *37*(8), 1077-1087. doi: 10.3758/MC.37.8.1077

Morey, R. D., & Rouder, J. N. (2018). BayesFactor: Computation of Bayes Factors for Common Designs [Computer software]. Retrieved from `https://CRAN.R-project.org/package=BayesFactor` (Version 0.9.12-4.2.)

Müller, K. (2017). here: A Simpler Way to Find Your Files [Computer software]. Retrieved from `https://CRAN.R-project.org/package=here` (Version 0.1)

Nunes, L. D., & Weinstein, Y. (2012). Testing improves true recall and protects against the build-up of proactive interference without increasing false recall. *Memory*, *20*(2), 138-154. doi: 10.1080/09658211.2011.648198

Pastötter, B., & Bäuml, K.-H. T. (2014). Retrieval practice enhances new learning: The forward effect of testing. *Frontiers in Psychology*, *5*. doi: 10.3389/fpsyg.2014.00286

Pastötter, B., Schicker, S., Niedernhuber, J., & Bäuml, K.-H. T. (2011). Retrieval during learning facilitates subsequent memory encoding. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *37*(2), 287-297. doi: 10.1037/a0021801

Pyc, M. A., & Rawson, K. A. (2010). Why Testing Improves Memory: Mediator Effectiveness Hypothesis. *Science*, *330*(6002), 335-335. doi: 10.1126/science.1191465

R Core Team. (2019). *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing. Vienna, Austria. (Version 3.5.3)

Re, A. C. D. (2013). compute.es: Compute Effect Sizes [Computer software]. Retrieved from `https://CRAN.R-project.org/package=compute.es` (Version 0.2-2)

Revelle, W. (2018). *Psych: Procedures for Psychological, Psychometric, and Personality Research*. Evanston, Illinois: Northwestern University. (Version 1.8.12)

Roediger, H. L., & Marsh, E. J. (2005). The Positive and Negative Consequences of Multiple-Choice Testing. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31*(5), 1155-1159. doi: 10.1037/0278-7393.31.5.1155

Roediger III, H. L., & Butler, A. C. (2011). The critical role of retrieval practice in long-term retention. *Trends in Cognitive Sciences*, *15*(1), 20-27. doi: 10.1016/j.tics.2010.09.003

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

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

Rowland, C. A. (2014). The effect of testing versus restudy on retention: A meta-analytic review of the testing effect. *Psychological Bulletin*, *140*(6), 1432-1463. doi: 10.1037/a0037559

Sahakyan, L., & Kelley, C. M. (2002). A contextual change account of the directed forgetting effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *28*(6), 1064-1072. doi: 10.1037/0278-7393.28.6.1064

Smith, T. A., & Kimball, D. R. (2010). Learning from feedback: Spacing and the delay–retention effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *36*(1), 80-95. doi: 10.1037/a0017407

Soderstrom, N. C., & Bjork, R. A. (2014). Testing facilitates the regulation of subsequent study time. *Journal of Memory and Language*, *73*, 99-115. doi: 10.1016/j.jml.2014.03.003

Szpunar, K. K., Khan, N. Y., & Schacter, D. L. (2013). Interpolated memory tests reduce mind wandering and improve learning of online lectures. *Proceedings of the National Academy of Sciences*, *110*(16), 6313-6317. doi: 10.1073/pnas.1221764110

Szpunar, K. K., McDermott, K. B., & Roediger, H. L. (2007). Expectation of a final cumulative test enhances long-term retention. *Memory & Cognition*, *35*(5), 1007-1013. doi: 10.3758/BF03193473

Szpunar, K. K., McDermott, K. B., & Roediger, H. L. (2008). Testing during study insulates against the buildup of proactive interference. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *34*(6), 1392-1399. doi: 10.1037/a0013082

Szpunar, K. K., Moulton, S. T., & Schacter, D. L. (2013). Mind wandering and education: From the classroom to online learning. *Frontiers in Psychology*, *4*. doi: 10.3389/fpsyg.2013.00495

Tabachnick, B. G., & Fidell, L. S. (2012). *Using Multivariate Statistics.* London, UK: Pearson.

Tulving, E., & Watkins, M. J. (1974). On Negative Transfer: Effects of Testing One List on the Recall of Another. *Journal of Verbal Learning and Verbal Behavior*(13), 181-193. doi: https://doi.org/10.1016/S0022-5371(74)80043-5

Wahlheim, C. N. (2015). Testing can counteract proactive interference by integrating competing information. *Memory & Cognition*, *43*(1), 27-38. doi: 10.3758/s13421-014-0455-5

Weinstein, Y., Gilmore, A. W., Szpunar, K. K., & McDermott, K. B. (2014). The role of test expectancy in the build-up of proactive interference in long-term memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *40*(4), 1039-1048. doi: 10.1037/a0036164

Weinstein, Y., McDermott, K. B., & Szpunar, K. K. (2011). Testing protects against proactive interference in face–name learning. *Psychonomic Bulletin & Review*, *18*(3), 518-523. doi: 10.3758/s13423-011-0085-x

Weinstein, Y., McDermott, K. B., Szpunar, K. K., Bäuml, K.-H., & Pastötter, B. (2015). *Not All Retrieval During Learning Facilitates Subsequent Memory Encoding.* Presented at the Annual Meeting of the Psychonomic Society. Chicago, IL.

Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis.* New York, NY: Springer-Verlag. (Version 3.1.0)

Wickham, H. (2017). tidyverse: Easily Install and Load the 'Tidyverse' [Computer software]. Retrieved from `https://CRAN.R-project.org/package=tidyverse` (Version 1.2.1)

Wickham, H. (2018). conflicted: An Alternative Conflict Resolution Strategy [Computer software]. Retrieved from `https://CRAN.R-project.org/package=conflicted` (Version

1.0.1)

Wissman, K. T., & Rawson, K. A. (2018). Test-potentiated learning: Three independent replications, a disconfirmed hypothesis, and an unexpected boundary condition. *Memory*, *26*(4), 406-414. doi: 10.1080/09658211.2017.1350717

Wissman, K. T., Rawson, K. A., & Pyc, M. A. (2011). The interim test effect: Testing prior material can facilitate the learning of new material. *Psychonomic Bulletin & Review*, *18*(6), 1140-1147. doi: 10.3758/s13423-011-0140-7

Xie, Y. (2019). knitr: A General-Purpose Package for Dynamic Report Generation in R [Computer software]. Retrieved from `https://CRAN.R-project.org/package=knitr` (Version 1.12)

Yang, C., Potts, R., & Shanks, D. R. (2018). Enhancing learning and retrieval of new information: A review of the forward testing effect. *NPJ science of learning*, *3*(1), 8. doi: 10.1038/s41539-018-0024-y

Zaromb, F. M., & Roediger, H. L. (2010). The testing effect in free recall is associated with enhanced organizational processes. *Memory & Cognition*, *38*(8), 995-1008. doi: 10.3758/MC.38.8.995

Zhu, H. (2019). kableExtra: Construct Complex Table with 'kable' and Pipe Syntax [Computer software]. Retrieved from `https://CRAN.R-project.org/package=kableExtra` (Version 1.1.0)