

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

Matej Pavlič^a, Denis Vlašiček^a & Dragutin Ivanec^a

a: Faculty of Humanities and Social Sciences, Department of Psychology, University of
Zagreb, Croatia

Corresponding author:

Denis Vlašiček

Ivana Lučića 3, 10000 Zagreb, Croatia

dvlasice@ffzg.hr

Total word count: 4625

Introduction + discussion word count: 2197

Abstract

This study explored the effects of retrieval and feedback on test-potentiated new learning. Participants read a text divided into three parts, between which they engaged in either episodic retrieval, semantic retrieval, or rereading. Participants in the retrieval conditions were randomly assigned to either receive or not to receive feedback on their achievement. We administered multiple choice questions whose distractors were designed specifically to facilitate 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. Feedback was found to have no bearing on new learning — neither on its own, nor via interaction with the interpolated activity type. No effect regarding the number of proactive intrusions was found, although exploratory Bayesian analyses preclude rejecting an effect. 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

General Audience Summary

Contemporary cognitive science has shown that activities people engage in after a study session have a significant effect on the learning outcome. The choice of whether to simply reread, or to rather attempt to recall the studied information from memory, has a non-negligible impact on long-term retention, with studies overwhelmingly supporting the latter option. Furthermore, trying but failing to retrieve a piece of information may have a silver lining in that it can actually make it more susceptible to successful memory processing in the future. It is assumed that active retrieval engages and refines generic mental processes that are also called upon at a more critical hour — during some final examination.

A closely related line of research seems to suggest that a memory test can even boost future learning of unrelated information. This somewhat puzzling effect is called test-potentiated *new* learning. Moreover, a number of studies examined a possibility that the key ingredient is not necessarily a test pertaining to the studied information. It was suggested that retrieval in general, be it from short-term or long-term memory, could be the cause of improvement in future learning. In short, retrieval itself seems to enhance memory for events that follow.

However, these studies mostly used fairly simple learning materials such as word lists. Our study, therefore, examined whether this effect generalises to everyday circumstances. Three groups of participants studied three paragraphs of text in sequence, and engaged in different activities between studying. One group reread the texts, another was given general knowledge questions, and the last group was given questions concerning the paragraphs. Our results show that when it comes to learning information that is typically found in an educational context, the safest bet would be to insert memory tests related to the studied material in between your study sessions.

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

The term “testing effect” refers to the finding that, when it comes to long-term retention of a piece of information, retrieving it 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). Besides directly enhancing retention through repetition of *successful* retrieval, testing effects can be brought about indirectly (Arnold & McDermott, 2013b; Roediger III & Karpicke, 2006a; but for a different view, see Kornell, Klein, & Rawson, 2015). For example, *unsuccessful* retrieval attempts can, through subsequent repeated encoding, generate test-potentiated (re)learning (TPL), whereby taking more tests increases the proportion of *newly* retrieved items in a test immediately following a restudy episode (Arnold & McDermott, 2013a, 2013b; Izawa, 1966, 1970; Kornell, Hays, & Bjork, 2009; Wissman & Rawson, 2018).

More importantly, following an initial impetus provided by Szpunar, McDermott, and Roediger (2008), who built upon earlier findings (Darley & Murdock, 1971; Tulving & Watkins, 1974), a decade of research has shown that retrieving previously studied information can even facilitate the acquisition of *new* information (Chan, Meissner, & Davis, 2018; Pastötter & Bäuml, 2014; Yang, Potts, & Shanks, 2018). If, within a multi-list learning paradigm, each subsequent study episode contains new items, testing the memory for those items after each learning episode still yields a greater number of correct responses and a decrease in proactive interference (PI) on a test pertaining to the final learning episode (e.g. Szpunar, Khan, & Schacter, 2013; Szpunar et al., 2008; Wissman, Rawson, & Pyc, 2011). Following the reasoning of Chan, Meissner, and Davis (2018), we use the term “test-potentiated *new* learning” (TPNL) to denote this effect. Having been investigated mainly using simple materials, a particularly important question for real-world applications is whether TPNL generalises to materials more complex than single words and word pairs. The research conducted in the preceding decade mostly points to a positive answer.

Wissman et al. (2011) were the first to ask whether the scope of TPNL extends to prose passages, and they did find a consistent effect using a free recall measure throughout their five experiments. As for its possible cause, evidence for an improvement via integration of

related information or via PI reduction — an explanation which figures prominently in the multi-list paradigm literature (Darley & Murdock, 1971; Szpunar et al., 2008) — was rather unconvincing. The effect was observed regardless of the interrelatedness of passages, and, possibly owing to the usage of text passages and free recall, overall intrusion rates were quite low. The authors subscribe to a view that retrieval attempts may promote the use of more effective encoding strategies.

In a similar study, Divis and Benjamin (2014) hypothesised that retrieval attempts enhance context fluctuations, which may lead to TPNL. Context changes have mostly been studied in terms of their role in *directed forgetting*, where it has been proposed that they could reduce interference (e.g. Sahakyan & Kelley, 2002). Since people store contextual alongside the target information during learning (Howard & Kahana, 2002; Mensink & Raaijmakers, 1988; Tulving & Thomson, 1973), the interplay between the two may affect memory performance. Regarding TPNL, it has been proposed that testing causes an internal context change relative to the one during study (Jang & Huber, 2008; Jonker, Seli, & MacLeod, 2013; Pastötter, Schicker, Niedernhuber, & Bäuml, 2011). Hence, items from successive study episodes might be updated with unique contextual information related to retrieval (Karpicke, Lehman, & Aue, 2014; Lehman, Smith, & Karpicke, 2014), while those encountered last remain associated only with the study context. For example, while knowing that each study episode contains new items, being aware that a recalled item was actually retrieved on a prior test could prevent subjects from providing erroneous responses, which may underlie the PI reduction observed in TPNL (Johnson, Hashtroudi, & Lindsay, 1993; Szpunar et al., 2008). In other words, additional contextual cues increase the disparity between pre- and post-retrieval items, providing an advantage to the latter by constraining the memory search set to items that are associated exclusively with the study context (Szpunar et al., 2008). However, if this scenario were true, then perhaps *any* interpolated task causing an internal context change might also produce TPNL.

Nonepisodic recall

One of the more curious findings in the field is that TPNL may indeed ensue not only after retrieving the previously studied material (episodic retrieval), but also after the retrieval of information unrelated to the studied material from semantic (Divis & Benjamin, 2014; Pastötter et al., 2011), or even 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 only the three forms of retrieval induced TPNL. In their first experiment, Divis and Benjamin (2014) adapted the procedure from Pastötter et al. (2011). Participants learned five lists of 10 words, between which they either retrieved information from semantic memory (e.g. enumerated professions) or counted backwards by 3s from a three-digit number. Only semantic retrieval enhanced performance in final list recall. They replicated and extended these findings in their second experiment by using complex learning materials. Participants read four text passages in a self-paced manner, while engaging in one of the same two types of intervening tasks between reading.

The argument these two groups of authors invoke to explain their results is that nonepisodic retrieval tasks sufficiently alter participant's internal context, and 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. An additional prediction stemming from the context change hypothesis is that introducing a delay between the last study episode and the final test should annul the benefits conferred by the contextual segregation because the delay offers ample opportunity for a context change even without the retrieval (Chan, Meissner, & Davis, 2018). Divis and Benjamin (2014) confirmed this in their final experiment, but Chan, Manley, Davis, and Szpunar (2018) provide evidence to the contrary, and side with the strategy change explanation suggested by Wissman et al. (2011). Chan, Manley, et

al. (2018), however, did not include any other supposed context-changing interpolated tasks besides episodic retrieval, which prevented assessing other predictions of the hypothesis.

Although Divis and Benjamin (2014) did find a reduction in PI in their first experiment, which they had predicted based on the purported isolating effect of context fluctuations, when their methodology precluded such effects in their second experiment, the results, resonating with those of Wissman et al. (2011), still showed TPNL. They conceded that PI reduction may not be the sole basis for the observed enhanced performance, and mention optimisation of encoding strategy as a viable cause. Finally, it is worth noting that, apart from prose passages, positive evidence for TPNL with complex learning materials was also found using video lectures (Jing, Szpunar, & Schacter, 2016; Szpunar, Khan, & Schacter, 2013), but in order to clarify the nature of TPNL within realistic settings, further inquiry is warranted.

Feedback

Even though it is not necessary for the testing effect to emerge (Roediger III & Karpicke, 2006b), corrective feedback is known to augment it (Roediger III & Butler, 2011). When it is provided, corrective feedback increases learning because it promotes error correction (Pashler, Cepeda, Wixted, & Rohrer, 2005) and reassurance (Butler, Karpicke, & Roediger, 2008). Feedback is particularly important for recognition tests such as multiple-choice tests since the usual benefit testing provides might turn into a disadvantage in case the test-taker selects a lure (Marsh, Roediger, Bjork, & Bjork, 2007; Roediger III & Marsh, 2005). Further, evidence points to the timing of feedback being a relevant variable when gauging its influence on learning, with delayed feedback given in bulk showing superior effects compared to immediate, piecemeal feedback (Butler, Karpicke, & Roediger, 2007; Butler & Roediger, 2008; Metcalfe, Kornell, & Finn, 2009; Smith & Kimball, 2010).

Perhaps unsurprisingly, there is a paucity of research into the potential role of feedback in TPNL because the process by which it could affect new learning is less apparent. One possibility is that feedback may act indirectly by guiding future efforts and motivating a change in encoding strategy (Roediger III & Karpicke, 2006a). Indeed, even failed or unconfident retrieval can be conceptualised as feedback per se because it can provide insight

into where additional cognitive resources should be allocated to improve performance (see e.g., Kornell et al., 2009). Moreover, providing feedback may engender additional processing (Bangert-Drowns, Kulik, Kulik, & Morgan, 1991) which could increase the likelihood of intrusions during new learning (Chan, Meissner, & Davis, 2018). Thus, the variable of corrective feedback may be a fruitful avenue for research.

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. Even though it has been shown that, in the standard multi-list procedure, substantially larger effect sizes follow after using free recall rather than recognition-level retrieval (Chan, Meissner, & Davis, 2018), choosing to examine the impact of feedback on TPNL imposed constraints upon our choice of testing format; immediate provision of feedback would have been intractable had we chosen to use free recall. Recognition is thought to be based more on familiarity than on controlled memory search (Yonelinas, 2002), which could be the reason behind the discrepant effect sizes. We, therefore, made an effort to heed the advice provided by Chan and McDermott (2007) — constraining the time for encoding, increasing the similarity between the correct answer and the distractors, and providing subjects with ample time to respond to questions — in order to encourage the use of controlled recollection.

Even though the relevant studies found PI to be superfluous when complex learning materials are used, in order to explore this possibility more closely, we used multiple-choice questions designed to assess memory both in terms of correct answers and susceptibility to intrusions. Secondly, there is a relative dearth of investigations using nonepisodic retrieval and recognition, and furthermore a lack of studies introducing corrective feedback (Chan, Meissner, & Davis, 2018). We therefore formed two memory tests, one of which tapped into episodic (assessing memory of the studied materials) while the other tapped into semantic memory (assessing general knowledge). Participants either were or were not given feedback upon completing an interpolated activity episode.

Based on the preceding discussion, if TPNL is mediated by context fluctuations, then

we should find that both types of retrieval enhance new learning, whereas rereading, which is deemed unable to bring about context changes (Chan, Meissner, & Davis, 2018), does not. If PI is in fact nonessential, then we should observe the expected memory boost without necessarily seeing an accompanying decrease in PI¹. We assumed that presenting feedback would have a positive effect on memory performance, but only for the participants engaging in episodic recall because only they could alter their encoding strategies accordingly. 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, 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 an 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,

¹Our original hypothesis suggested a differential effect of activity type on PI rates. Specifically, we expected that participants engaging in episodic retrieval would display the lowest susceptibility to PI, followed by participants in the semantic retrieval condition, and finally by those tasked with rereading. Upon further examination of the literature, our expectations changed, but this had no bearing on the analyses, which remained the same nevertheless.

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 explicitly 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. Unbeknownst to them, 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. Still, the lowest possible time limit was set to 5, and the longest to 8 minutes.

Next, participants were familiarised with the interpolated activity they were going to

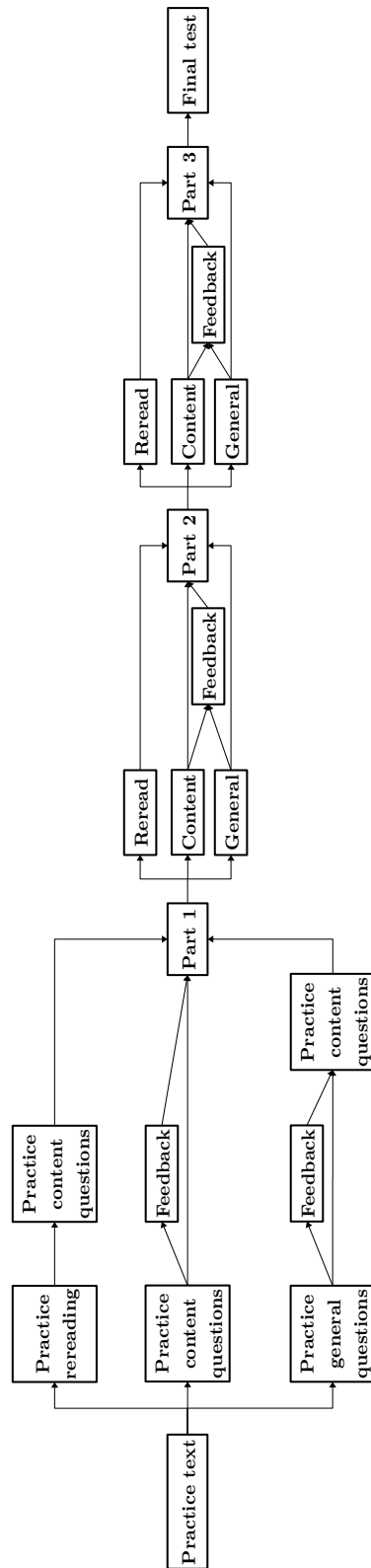


Figure 1. A flowchart depicting the experimental procedure.

perform during the main part of the procedure. The content-related test group answered 4 questions based on the practice text, the general-knowledge test group answered 4 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 4 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. All tests were self-paced and no time limit was applied. 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. Depending on the condition they were assigned to, they could also receive 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 20 novel questions examining their knowledge of that part. Through this increase of the number of questions we sought to avoid a ceiling effect on our main dependent variable, which could mask a real effect of TPNL. No feedback was presented after the final test, irrespective of the experimental condition.

In total, 44 content related questions with 4 response options were generated from the presented parts of the text. 4 questions were presented after the practice text, 10 after each of the first two parts (only to the participants in the content related test condition), and 20 after the third part of the text (to all participants). Starting from the second test, the distractor options were chosen so that (i) two distractors were plausible, but unrelated to the text, and (ii) 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 “intruder” in the rest of this article). For example, a question in the final test was *The agricultural habitat on which a plant grows is called the...* The response options were (a) agrobiosphere (distractor, never mentioned in the text), (b) agrosphere (intrusive distractor, mentioned in the previous part of the text), (c) biotope (distractor), (d) agrobiotope (correct

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

answer). Further, 24 general knowledge questions were generated. These were presented to participants in the general-knowledge test condition, after the first two parts of the text and after the practice text. The questions spanned recent history, popular culture, and art (e.g. *Kurt Vonnegut's famous anti-war novel is called...*).

The computer recorded whether a participants response was correct and whether they 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.

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.

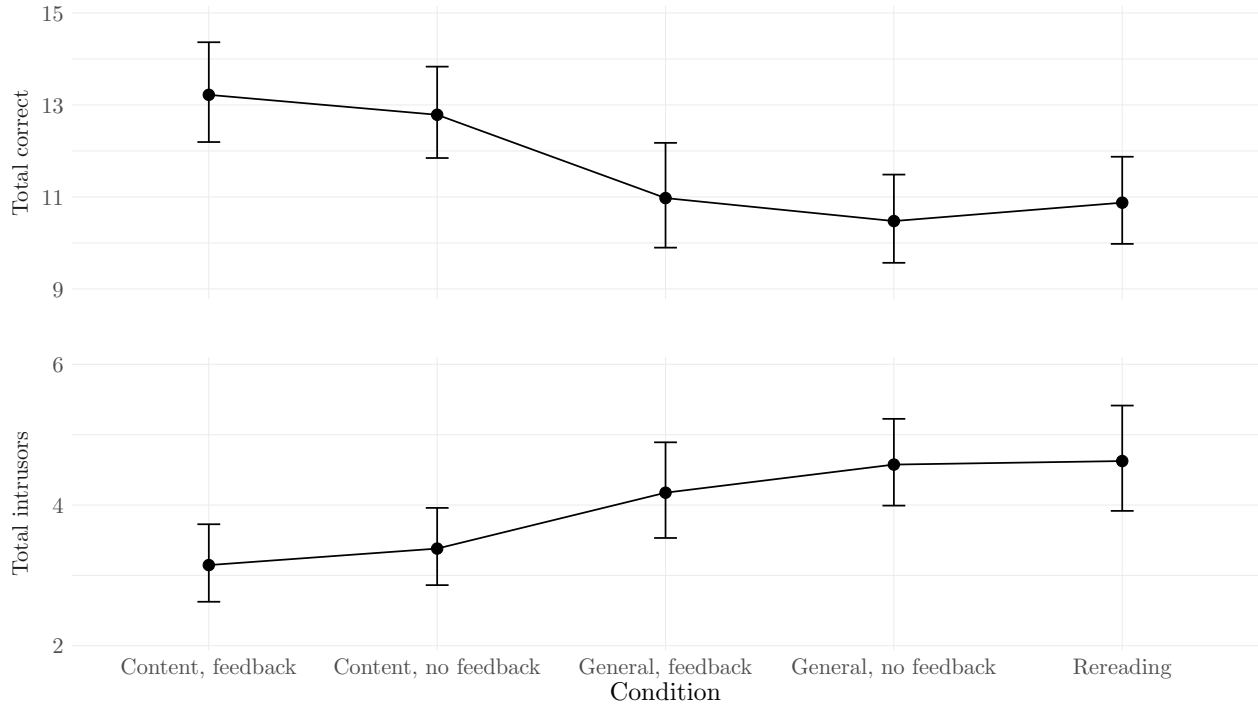


Figure 2. Mean number of correct answers and mean number of intrusive distractors chosen, broken down by experimental condition. Error bars show the 95% confidence intervals around the means.

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.

The correlation between our DVs calculated on the whole sample is $r(201) = -.707$ (95% CI: $[-.77, -.63]$, $p < .0001$). Given that we have two dependent variables, which are highly correlated, we have decided to conduct a one-way MANOVA. According to [Tabachnick and Fidell \(2012\)](#), conducting a MANOVA instead of multiple ANOVAs increases the chance of discovering the effects of different treatments. Furthermore conducting a MANOVA guards against the inflation of Type 1 errors due to multiple tests of correlated dependent variables ([Field, Miles, & Field, 2012](#); [Tabachnick & Fidell, 2012](#)). Finally, conducting

separate ANOVAs would disregard the correlation between our two dependent variables (Field et al., 2012). Therefore, 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.

A power analysis conducted prior to analyzing the data (using the G*Power software by Faul, Erdfelder, Buchner, & Lang, 2009) has shown that we should have above 80% power to detect effects which fall between small and medium (Cohen's $f^2 \gtrsim 0.06$), at an α level of .025, with a sample size of 110 participants. Note that larger effects are expected based on prior studies.

Pillai's V for the analysis is .126, $p = .004$ (Wilks' $\Lambda = .875$, $p = .003$). The effect size, calculated as $\omega_{mult}^2 = .109$ (bootstrap median² = .132, BC_α 95% CI = [.011, .202])³. 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). 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 a differential effect of our conditions on the total number of correct answers, 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 type on the total number of chosen intrusors ($F(2, 118) = 0.844$, $p = .433$). So far, our analyses show a lack of evidence in support of our second hypothesis that the type of

²All bootstrap estimates taken from 10000 replications.

³Cohen's $f^2 = 0.051$ (calculated according to Equation 12 in Steyn Jr & Ellis, 2009).

interpolated activity will have an effect on the number of intrusors.

In order to test our first hypothesis, we contrasted (i) the two test groups with each other, and (ii) the rereading group with the two test groups, taking only the total number of correct answers as the DV. We found 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%). Students in the content-related-test condition score higher on the final test than students in the general-knowledge-test condition. The second contrast shows 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%). Therefore, we cannot conclude that being in the rereading condition, as opposed to being in one of the two test groups, leads to different learning outcomes. However, we believe this is due to the rereading and general-knowledge-testing groups having very similar mean numbers of correct answers ($M = 10.88$ and 10.47 , respectively) 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. Again, a power analysis conducted before analysing the data has shown that we should have above 80% power to detect effects which fall between small and medium (Cohen's $f^2 \gtrsim 0.05$), at an α level of .025, with a sample size of 145 participants.

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 $\omega_{mult}^2 = .065$ (bootstrap median = .072, BC_α 95% CI = [.008,

.140])⁴. 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 estimated effect size is $\omega_{mult}^2 = 0$. 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 estimated effect size $\omega_{mult}^2 = 0$.

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 students in the content-related-test condition score higher on the final test than students 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 ($F(1, 155) = 0.175$, $p = .676$). Therefore, we find no evidence for an effect of the type of interpolated activity nor of feedback on the number of intrusive distractors chosen. 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.

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.

⁴Cohen’s $f^2 = 0.063$, using Equation 12 from [Steyn Jr and Ellis \(2009\)](#).

Table 2

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

Term	<i>SS</i>	<i>df</i>	<i>F</i>	<i>p</i>
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		

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. 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 showed that observed differences were driven primarily by the number of correct responses, while finding no evidence for the contribution of PI.

The fact that we observed the effect of interest while employing a testing format which

is known to produce the smallest effects is interesting in and of itself, and suggests that the effect should hold in conditions that are arguably the most prevalent in western educational systems. Nevertheless, our results are not entirely in line with extant research. While our results point to an exclusive role of episodic retrieval, [Pastötter et al. \(2011\)](#), for example, found that both types of retrieval can generate TPNL. Notably, these authors used simpler learning materials and free recall — both learning material complexity and testing format are known moderators of TPNL ([Chan, Meissner, & Davis, 2018](#)). However, the few studies that examined the effects of nonepisodic recall on TPNL have produced equivocal results. While two studies suggested that nonepisodic and episodic recall have comparable effects ([Divis & Benjamin, 2014](#); [Pastötter et al., 2011](#)), [Weinstein et al. \(2015\)](#) failed to show this. Among a number of methodological differences between these studies, the specific type of nonepisodic recall stands out as a possible reason behind the diverging results. While [Pastötter et al. \(2011\)](#) and [Divis and Benjamin \(2014\)](#) both used semantic generation, [Weinstein et al. \(2015\)](#) used recall from autobiographical memory. Delineating the potential distinctive effects various forms of non-episodic recall could have in the TPNL paradigm is a goal future studies may pursue.

Studies that have suggested that nonepisodic recall may serve as an effective method of learning potentiation have drawn on context theories to explain their results ([Divis & Benjamin, 2014](#); [Pastötter et al., 2011](#)). [Divis and Benjamin \(2014\)](#) proposed that retrieval processes enhance context fluctuation, thereby increasing the contextual disparity between information acquired across study sessions. This, in turn, reduces the memory search set and PI. The absence of an effect of semantic retrieval on learning in our study may be taken as evidence against this account of TPNL because, presumably, semantic retrieval should have produced the internal context change required for resetting the encoding process ([Pastötter et al., 2011](#)). A Bayesian estimate of the effect of nonepisodic recall also lends support for the claim that it does not enhance learning (see supplementary materials). Still, a basic assumption we have made is that the interpolated activity that served the function of activating retrieval from semantic memory in our study was effective.

While we found no evidence for an effect of feedback on TPNL, exploratory Bayesian

analyses do not exclude the possibility of a feedback effect, but the obtained estimates point to an effect which could be practically equivalent to zero. From this we gather that the collected data provide no evidence that a PI reduction mechanism underpins TPNL. Interpreting these results warrants caution, though, since a more precise estimate of the effect is desirable. Assuming that the effect of feedback was in fact insignificant, this may have occurred because episodic recall in and of itself gave subject the information they required in order to adjust their encoding strategy (Kornell et al., 2009). If this were the case, additional explicit feedback might have been wholly redundant.

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 PI. However, previous work has shown that release from PI may play basically no role when it comes to learning complex materials (Divis & Benjamin, 2014; Wissman et al., 2011). The recognition-level method which we employed showed no signs of PI beyond those expected to occur by chance alone. Admittedly, limiting the number of choices by displaying possible answers may have diluted the interference effect other unwritten pieces of information might have had, if we had used free recall instead.

Recently, Chan, Meissner, and Davis (2018) provided a meta-analytic analysis and comprehensive overview of the literature, identifying four theoretical frameworks that figure prominently as explanations for TPNL. Besides context theories, whose main points are outlined above, *resource theories* generally posit that testing increases cognitive resources either by (i) reducing encoding-based PI (e.g. Nunes & Weinstein, 2012; Szpunar et al., 2008; Wahlheim, 2015; Weinstein, McDermott, & Szpunar, 2011), (ii) restoring encoding/attentional resources (e.g. Pastötter et al., 2011), or by (iii) altering mind wandering patterns to focus them more on the learning material (e.g. Jing et al., 2016; Szpunar, Khan, & Schacter, 2013; Szpunar, Moulton, & Schacter, 2013). A circumscription of separate learning episodes is at the core of both resource and context accounts, but its effect on learning is different depending on the outlook. According to the former, isolating a learning episode through attempts at recall increases resources for subsequent learning by preventing *encoding-based* PI. On the other hand, context theories emphasise later *retrieval* processes, whereby isolating

an earlier learning episode reduces the memory search set for retrieval.

While resource theories revolve around the amount of deployable cognitive resources, *metacognitive theories* emphasise the optimisation of their deployment, regardless of their amount. Thus, using metacognitive knowledge gained through retrieval attempts (e.g. [Chan, Manley, et al., 2018](#); [Cho, Neely, Crocco, & Vitrano, 2017](#)), and perhaps even failures of retrieval ([Bahrick & Hall, 2005](#)), learners can adjust their strategies. In line with these proposals, recent studies have shown that 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](#)). For example, in a recent investigation using multi-list learning, [Chan, Manley, et al. \(2018\)](#) found that, compared to untested groups, the group subjected to interpolated testing displayed an increase in category-based clustering during free recall. They ascribe this to an adoption of more efficient encoding/retrieval strategies. A similar pattern is often observed in studies exploring the testing effect, where a greater number of tests is associated with improved organisation of output displayed upon testing ([Karpicke, 2012](#); [Zaromb & Roediger, 2010](#)).

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 prior-learning material. On one account, testing increases the likelihood of spontaneous covert retrieval of older 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.

The accounts mentioned above are not mutually exclusive. For example, one can imagine how a strategy shift could reduce PI and foster integration. Still, we believe our results are mostly aligned with a metacognitive explanation of the effect ([Chan, Manley, et al., 2018](#); [Wissman et al., 2011](#)).

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 following the last reading episode. The typical instruction given to participants in the TPNL paradigm is that interpolated activities will be determined randomly (Yang et al., 2018). Thus, an attempt is made to equalise 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 procedure, regardless of the instructions they are given (Weinstein, Gilmore, Szpunar, & McDermott, 2014). If they take a test, they will more likely expect another one, and such expectations are known to influence encoding (e.g. Szpunar, McDermott, & Roediger, 2007).

Further, we cannot exclude the possibility that our interpolated activities had differential effects on our participants' motivation. Several participants remarked 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 should have observed the lowest scores in the rereading group. Yet, this is not the case. Furthermore, we find no reason to believe that answering general knowledge questions is less interesting than answering questions about weeds, and presume that the former would, therefore, help sustain motivation. On the other hand, differences in engagement 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 had been a TPNL effect, we would expect higher scores on the second test. This points to the tentative conclusion that the interpolated tests themselves differed in difficulty. Thus, we cannot reject the possibility that differing difficulties affected 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 PI reduction accounts within the wider framework of resource theories.

Author Contributions

All three authors participated in the experimental design and design of the materials. MP programmed the application used for data collection. MP and DV collected the data and analysed it. MP and DV wrote the manuscript, DI served as editor.

Acknowledgements

We would like to thank every one of our participants for making this study possible. Also, we would like to thank Marijana Glavica, our librarian, who has been tremendously helpful around data archiving and preparations for preprint publication.

Notes

Analyses conducted using the *R* language (R Core Team, 2019). 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).

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 <https://github.com/ffzg-erudito/inter-testing-feedback-2018>. Materials are also available through <https://osf.io/gk9a3/>. The data is also hosted on <https://dataverse.ffzg.unizg.hr/dataset.xhtml?persistentId=doi:10.23669/JVNVNR>.

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
- Bangert-Drowns, R. L., Kulik, C.-L. C., Kulik, J. A., & Morgan, M. (1991). The Instructional Effect of Feedback in Test-like Events. *Review of Educational Research*, 61(2), 213–238. doi: 10.2307/1170535
- 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., Karpicke, J. D., & Roediger, H. L. (2008). Correcting a metacognitive error: Feedback increases retention of low-confidence correct responses. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 34(4), 918–928. doi: 10.1037/0278-7393.34.4.918
- 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., & McDermott, K. B. (2007). The testing effect in recognition memory: A dual process account. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 33(2), 431–437. doi: 10.1037/0278-7393.33.2.431
- 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
- Faul, F., Erdfelder, E., Buchner, A., & Lang, A.-G. (2009). Statistical power analyses using G* Power 3.1: Tests for correlation and regression analyses. *Behavior research methods*, *41*(4), 1149–1160.
- Field, A., Miles, J., & Field, Z. (2012). *Discovering Statistics Using R*. Thousand Oaks, CA: SAGE Publications Ltd.
- 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.
- Howard, M. W., & Kahana, M. J. (2002). A distributed representation of temporal context. *Journal of Mathematical Psychology*, 268–299.
- 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
- Johnson, M. K., Hashtroudi, S., & Lindsay, D. S. (1993). Source Monitoring. *Psychological Bulletin*, 114(1), 3–28. doi: <https://doi.org/10.1037/0033-2909.114.1.3>
- Jonker, T. R., Seli, P., & MacLeod, C. M. (2013). Putting retrieval-induced forgetting in context: An inhibition-free, context-based account. *Psychological Review*, 120(4), 852–872. doi: 10.1037/a0034246
- 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., Lehman, M., & Aue, W. R. (2014). Retrieval-Based Learning. In *Psychology of Learning and Motivation* (Vol. 61, pp. 237–284). Elsevier. doi: 10.1016/B978-0-12-800283-4.00007-1
- 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
- Mensink, G.-j., & Raaijmakers, J. G. W. (1988). A model for interference and forgetting.

- Psychological Review*, 434–455.
- 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
- 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
- Pashler, H., Cepeda, N. J., Wixted, J. T., & Rohrer, D. (2005). When Does Feedback Facilitate Learning of Words? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 31(1), 3–8. doi: 10.1037/0278-7393.31.1.3
- 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 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
- Roediger III, 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
- 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
- Steyn Jr, H. S., & Ellis, S. M. (2009). Estimating an effect size in one-way multivariate analysis of variance (MANOVA). *Multivariate Behavioral Research*, *44*(1), 106–129.
- 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., & Thomson, D. M. (1973). Encoding specificity and retrieval processes in episodic memory. *Psychological Review*, 80(5), 352–373. doi: 10.1037/h0020071
- 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](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. (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
- Yonelinas, A. P. (2002, April). The Nature of Recollection and Familiarity: A Review of 30 Years of Research. *Journal of Memory and Language*, 46(3), 441–517. doi: 10.1006/jmla.2002.2864
- 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)