Memory

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/pmem20

How retrieval practice and semantic generation affect subsequently studied material: an analysis of item-level effects

Oliver Kliegl & Karl-Heinz T. Bäuml

To cite this article: Oliver Kliegl & Karl-Heinz T. Bäuml (2023) How retrieval practice and semantic generation affect subsequently studied material: an analysis of item-level effects, Memory, 31:1, 127-136, DOI: 10.1080/09658211.2022.2127770

To link to this article: https://doi.org/10.1080/09658211.2022.2127770

Published online: 25 Sep 2022.

MEMORY
2023, VOL. 31, NO. 1, 127–136
https://doi.org/10.1080/09658211.2022.2127770

How retrieval practice and semantic generation affect subsequently studied material: an analysis of item-level effects

Oliver Kliegl and Karl-Heinz T. Bäuml

Department of Experimental Psychology, Regensburg University, Regensburg, Germany

ABSTRACT
The forward testing effect (FTE) refers to the finding that retrieval practice of previously studied material can facilitate recall of newly studied (critical) material. Such interim retrieval practice can also lead to a differential FTE, i.e., a more pronounced FTE for items at early than later serial positions in the critical material. The present study examined whether this differential FTE also holds with interim semantic generation of extra-list items, and whether it is influenced by study material. Consistent with prior work, the results of two experiments showed that both interim retrieval practice and interim semantic generation induce the general (list-level) FTE when unrelated study lists are applied, whereas retrieval practice only creates the effect with categorised study lists. Critically, however, the differential FTE was present in response to retrieval practice but absent in response to semantic generation. This pattern held regardless of which material was studied, thus experimentally dissociating the general (list-level) from the differential (item-level) FTE. The findings may suggest that retrieval practice, but not semantic generation, induces a reset of the encoding process which promotes attentional encoding such that a more pronounced FTE arises for early than middle and late serial positions in the critical list.

ARTICLE HISTORY
Received 2 June 2022
Accepted 16 September 2022

KEYWORDS
Recall; testing effect; retrieval practice; semantic generation; serial position effects

Both students and teachers generally view tests primarily as a means of measuring performance (Karpicke et al., 2009). However, tests can have tremendous benefits for the long-term retention of the practiced information. In fact, studies on this so-called testing effect have shown that retrieval practice of previously studied information often leads to improved retention of this information in a later test than, for example, restudy of the information does (e.g., Roediger & Karpicke, 2006). A myriad of laboratory studies has cemented the view that the testing effect is a robust and general phenomenon that occurs across a wide range of materials, age groups, and ability levels (for reviews, see Karpicke, 2017; Roediger & Butler, 2011).

However, retrieval practice improves not only retention of the practiced material itself, but also fosters the learning and memory of subsequently encountered, new material. For instance, Szpunar et al. (2008) asked subjects to study five word lists in succession in anticipation of a final cumulative retention test and to solve either simple mathematical tasks immediately after the presentation of lists 1–4 (distractor condition), study the word lists again (restudy condition), or attempt to retrieve the words from the immediately preceding list (retrieval-practice condition). After learning list 5, all subjects were asked to recall the words from this critical final list. The results showed that subjects in the retrieval-practice condition remembered more words from list 5 and showed fewer intrusions of words from lists 1–4 than subjects in the other two conditions. This effect, which is often referred to as the forward testing effect (FTE), is a very general effect that has been found in both lab-based studies and educational settings. It has been shown to arise across a variety of study materials, including word lists, paired associates (Weinstein et al., 2011), prose material (Wissman et al., 2011), and videos (Szpunar et al., 2013); and it has been shown to arise in a range of participant groups, like college students, children (Aslan & Bäuml, 2016), older adults (Pastötter & Bäuml, 2019), and individuals suffering from traumatic brain injury (Pastötter et al., 2013; for reviews, see Pastötter & Bäuml, 2014; Yang et al., 2018).

Two prominent FTE accounts

Two prominent accounts of the FTE are the context-change explanation and the strategy-change explanation. The context-change account is based on the widely held belief that during study of to-be-learned material, individuals also encode details about the mental context in which that information is provided (Estes, 1955; Mensink & Raaijmakers, 1988). Retrieval practice between the study of single lists is assumed to change this mental context and

CONTACT Oliver Kliegl oliver.kliegl@ur.de Department of Experimental Psychology, Regensburg University, 93040 Regensburg, Germany

thus create distinct temporal contexts for each study list, which enables a more focused memory search for the critical final list at the time of test (Bäuml & Kliegl, 2013; Szpunar et al., 2008). Evidence for the view, for instance, comes from research demonstrating that the FTE is accompanied by shorter response latencies for the critical final list (Bäuml & Kliegl, 2013; Lehman et al., 2014). Because shorter response latencies indicate a smaller memory search set (Wixted & Rohrer, 1993; Rohrer, 1996), the finding is consistent with the context-change account's assumption of a more effective discrimination between the prior lists and the critical list.

In contrast, the strategy-change explanation is based on the view that interim retrieval practice can lead subjects to consider new – and potentially more effective – strategies for further learning. Retrieval practice can indeed provide critical information about the learning task at hand, and may enable subjects to build expectations about the particular format of later tests or the presence of retrieval cues on these tests. On the basis of such information, encoding and/or retrieval strategies may be optimised (Chan, Manley, et al., 2018; Davis & Chan, 2015; Soderstrom & Bjork, 2014). Chan, Manley, et al. (2018), for instance, found that the FTE was accompanied by an improved semantic organisation of the critical list, as subjects in the retrieval-practice condition showed a stronger propensity to cluster their recall on the basis of the items' category membership. Together with the fact that the FTE can still arise when the retention interval prior to the test of the critical list or the lag prior to study of this list is prolonged, the finding is consistent with the strategy-change explanation. Critically, the context-change and strategy-change explanations are not mutually exclusive. In particular, a recent two-factor account of the FTE argues that both context change and strategy change can contribute to the FTE (Kliegl & Bäuml, 2021; see General Discussion).

An (additional) role of attentional factors for the FTE?

In recent years, also attentional accounts of the FTE have been proposed (e.g., Chan et al., 2020; Pastötter et al., 2011, 2018; Szpunar et al., 2013; Yang et al., 2017). Common to these accounts is the idea that, in the absence of any interim retrieval practice, attentional resources decline when several lists are studied in succession, assuming that it becomes more and more difficult to keep the attention focused if more and more material is to be encoded. Retrieval practice between study of the single lists may prevent such increase in inattention. Consistently, interim retrieval practice has been found to encourage task-relevant activities like note-taking and discourage task-irrelevant activities like mind wandering (Szpunar et al., 2013). Similarly, interim retrieval practice can prevent a reduction of study time across lists when encoding is self-paced (Yang et al., 2017).

A variant of these attentional accounts assumes that retrieval practice induces a reset of the encoding process that may reduce the need to keep the previously studied material in mind and help subjects to refocus their attentional resources (Pastötter et al., 2011). Such an encoding reset may arise because retrieval practice largely exhausts the opportunities to review studied material (e.g., Chan, Meissner, et al., 2018; Roets et al., 2006) – at least if the recall phase is sufficiently long. Indeed, unlike an interim restudy or distractor activity, interpolated retrieval practice may provide the subject with a relatively fair assessment of how well they are able to remember a list's items (for a similar suggestion, see Roediger, et al., 2011). Critically, even if the subject judges their retrieval performance to be suboptimal, in the absence of feedback, they cannot do much to improve the situation, so they may mentally close with the task in the latter part of the retrieval-practice phase. This proposal also fits with participants' subjective reports that, during the latter part of the recall phase, they are often "finished with the list" or feel "ready for the next task ahead". [1] Pastötter et al. (2018) recently showed that a more fine-grained analysis of recall data on the item level can provide critical insights about the nature of the attentional processes contributing to the FTE. These researchers had participants study three lists consisting of 12 unrelated items each. Results showed a typical FTE for the critical list, but analysis of the list items' serial position curves indicated that the size of the FTE varied with the items' serial position in the list. It was more pronounced for the list's early items at list primacy positions (items 1-4) than the list's middle and late items (items 5-12). These findings suggest that a retrieval-practice induced encoding reset may particularly benefit the encoding of the first few items of the study list presented after retrieval practice and thus induce a differential FTE.

The present study

The first goal of the present study therefore was to replicate the Pastötter et al. (2018) finding that retrieval practice can induce a differential FTE, with greater recall benefits for the early than the middle and late items of the critical list. The second goal was to examine the generalizability of the differential FTE. One factor that may play a critical role for the presence of the differential FTE is type of retrieval activity. Prior research has shown that the general (list-level) FTE is not only induced following retrieval practice of a previously studied list, i.e., an episodic type of test, but can result from various interim retrieval activities, including a semantic type of test in which subjects are asked to generate as many exemplars as possible from various extra-list semantic categories (e.g., *profession* or *four-legged animal*). Indeed, such semantic generation, when conducted between the study of single lists, has been found to create an FTE as well (Divis & Benjamin, 2014; Pastötter et al., 2011). Whether the differential FTE also transpires when interim semantic generation is applied instead of interim retrieval practice is unclear. However, if an encoding reset mediated the differential FTE, then the differential FTE should be restricted to retrieval practice and not generalise to semantic generation. Indeed, retrieval practice, but not semantic generation, may leave subjects with the feeling that they are "finished with the list", inducing an encoding reset.

Another factor that may influence the presence of the differential FTE is type of study material. Recent research in fact showed that the presence of the general FTE can depend on study material. After both prolonged retention interval between study and test of the critical list and prolonged lag prior to study of the critical list, Kliegl and Bäuml (2021) found the general FTE to disappear with unrelated, but not with categorised lists. Furthermore, the semantic-generation induced general FTE arose with unrelated, but not with categorised lists. Although type of material may thus be a potentially critical factor for FTE findings, there is reason to expect that it may not influence the differential FTE. Indeed, if an encoding reset was the driving force behind the differential FTE, then retrieval practice should induce such a reset regardless of whether the studied lists consist of unrelated or categorised material.

This study reports the results of two experiments designed to examine whether (i) the nature of the retrieval activity between study of the single lists and (ii) the type of study material affect whether a differential FTE arises. Separately for unrelated item lists (Experiment 1) and categorised item lists (Experiment 2), we examined whether the differential FTE is present after both interim retrieval practice and interim semantic generation. In both experiments, the three-list version of the typical FTE task was applied (Bäuml & Kliegl, 2013; Pastötter et al., 2018). After study of lists 1 and 2, subjects were either asked to study the immediately preceding list once again (restudy condition), recall the words of the immediately preceding list (retrieval-practice condition), solve simple arithmetic problems (distractor condition), or generate as many exemplars as possible from extra-list categories (semantic-generation condition). In all four conditions, study of the critical list 3 was followed by a free-recall test of the items of this list. For this test, mean recall rates and serial position curves were analyzed for each of the four study conditions. According to the encoding-reset hypothesis, the differential FTE should follow exclusively as a result of a preceding retrieval-practice, and should therefore not occur after preceding semantic generation of extra-list items. Moreover, such pattern of results should arise regardless of study material and thus be present in both Experiment 1 and Experiment 2.

Experiment 1

The goal of Experiment 1 was to examine whether, for unrelated study lists, retrieval practice but not semantic generation elicits a differential FTE – that is, a greater FTE for early than middle and late critical items – or whether both retrieval practice and semantic generation can produce such an effect. On the basis of prior work (Divis & Benjamin, 2014; Kliegl & Bäuml, 2021; Pastötter et al., 2011), a general (list-level) FTE was expected for this material both when there was retrieval practice of the previously studied lists (relative to the restudy baseline) and when there was semantic generation of extra-list items between study of the single lists (relative to the distractor baseline).

[2] We also anticipated that, for retrieval practice, the differential FTE would arise (Pastötter et al., 2018), likely reflecting an encoding reset. If an encoding reset was a necessary precondition for the differential FTE, then semantic generation should not induce the differential FTE.

Method

Participants

The required sample size in Experiments 1 and 2 was calculated using G*Power (Version 3.1.9.2; Faul et al., 2009). Given $\alpha = 0.05$ and a desired power of $1 - \beta = 0.85$ to detect the critical interaction between condition and serial position with a small to medium effect size of $f = .2$ or larger, a required sample of 65 subjects per experimental condition was calculated. Closely following this recommendation, we recruited 288 students at Regensburg University for Experiment 1 (mean age = 21.9 years), with 72 subjects in each of the four experimental conditions. Participants took part in the experiments in return for either partial course credit or a compensatory amount of money. All subjects spoke German as their native language. All reported experiments were carried out in accordance with the provisions of the World Medical Association Declaration of Helsinki. Participants in both experiments were tested in person. Data collection was finished before the COVID-19 outbreak.

Material

Seventy-two unrelated German nouns of medium frequency were drawn from the CELEX database (Duyck et al., 2004). For each participant, items were assigned randomly to three lists consisting of 24 items each. The study material was identical to the material applied in one of our earlier studies (Kliegl & Bäuml, 2021, Experiments 1a, 2a, 3a).

Design and procedure

Prior to the start of the experiment, participants were told that they would be asked to study several lists of items. They were also informed that they should anticipate various activities that may follow the presentation of each single list, which can include simple arithmetic tasks, restudy of a list that they had just previously studied, a free-recall test on all the words from a just studied list, or semantic generation of exemplars from a semantic category unrelated to the words that they had just studied. It was pretended that these interlist activities would occur on a completely random basis when, in fact, interlist activities differed between conditions. In particular, subjects engaged in the same interlist activities after the encoding of lists 1 and 2 within each experimental condition. Participants were also made aware that, regardless of these interlist activities, all study lists would be tested in a final cumulative test.

At the start of the experiment, the items of the three lists were visually presented at the centre of a computer screen, and the 24 words of each list were exposed individually for 4.5 s with a 0.5 s interitem interval. After the presentation of each single list, subjects counted backward in steps of threes from a random three-digit number for 30 s. Experimental conditions differed in the type of interlist activity that followed this backward counting after lists 1 and 2. Participants were either asked to (i) restudy the immediately preceding list (each item was again shown for 4.5 s per item with a 0.5 s interitem interval; restudy condition), (ii) were given 120 s to write down on a piece of paper as many words from the immediately preceding list as they could (retrieval-practice condition), (iii) solve simple arithmetic tasks for 120 s (distractor condition), or (iv) spend the same amount of time on a semantic-generation task (semantic-generation condition). In the semantic-generation task, subjects were given 60 s to write down as many German exemplars as possible from a first of four categories (four-legged animals, sports, vegetables, or professions), and then were given another 60 s to write down as many German exemplars as possible from a second of the four categories. Selection of categories after list 1 was random; after list 2, the remaining two categories were tested. After study of the critical list and the backward-counting task, participants in all conditions were asked to write down as many items as possible from the critical list on a piece of paper. They were given 120 s for this free-recall task. Following recall of the critical list, participants were also tested on lists 1 and 2. Again, participants had 120 s to write down as many of the list items as they could. Test order of lists 1 and 2 was random. Final-test performance of lists 1 and 2 is of no direct relevance for the present study and will not be reported.

Results

For all experiments, we provide Bayes factors (B_{01}) – which reflect the odds in favour of the null hypothesis over the alternative hypotheses – when a finding did not reach conventional level of statistical significance (i.e., $\alpha = .05$). For general orientation, a B_{01} ranging from 1–3 can be considered as anecdotal evidence for the null hypothesis, a B_{01} ranging from 3–10 as moderate evidence for the null hypothesis, and a B_{01} ranging from 10–30 as strong evidence for the null hypothesis (Masson, 2011; Raftery, 1995).

List-level analysis of critical-list recall

Correct recall

On the list level, the percentage of correctly recalled critical items is shown for each of the four experimental conditions in Figure 1a. An overall ANOVA of the four conditions (restudy, retrieval practice, distractor, semantic generation) showed a significant effect of CONDITION, $F(3,284) = 14.749$, $MSE = .035$, $p < .001$, partial $\eta^2 = .135$. Pairwise comparisons revealed that the difference of 15.6% between the restudy and retrieval practice condition was reliable, $F(1,144) = 35.283$, $MSE = .025$, $p < .001$, partial $\eta^2 = .199$, as was the difference of 13.4% between the distractor and semantic generation condition, $F(1,144) = 14.507$, $MSE = .046$, $p < .001$, partial $\eta^2 = .093$, thus demonstrating both retrieval-practice induced and semantic-generation induced FTEs. In contrast, the difference of 2.7% between the retrieval-practice and semantic-generation conditions was not significant, $F(1,144) < 1$, $B_{01} = 8.088$, as was the difference of 0.7% between the restudy and distractor conditions, $F(1,144) < 1$, $B_{01} = 11.675$.

Intrusions

All list-1 or list-2 items that participants produced during the recall test of the critical list were counted as intrusions. An overall ANOVA of the four conditions (restudy, retrieval practice, distractor, semantic generation) showed a significant effect of CONDITION, $F(3,284) = 7.852$, $MSE = .728$, $p < .001$, partial $\eta^2 = .077$. Pairwise comparisons revealed that neither the difference of 0.18 intrusions between the restudy and retrieval-practice condition were significant, $F(1,144) = 3.379$, $MSE = .347$, $p = .068$, partial $\eta^2 = .023$, $B_{01} = 2.206$, nor the difference of 0.16 intrusions between the distractor and semantic-generation condition, $F(1,144) < 1$, $B_{01} = 8.178$. The significant main effect of CONDITION primarily arose from a higher number of mean intrusions in both the distractor and semantic-generation conditions than either the restudy and retrieval practice conditions (see Table 1).

Item-level analysis of critical-list recall

On the item level, proportion of correct recall was analyzed as a function of CONDITION and the within-participants factor of SERIAL POSITION. To this end, the critical list was broken into eight bins spanning three items each (bin 1, items 1–3; bin 2, items 4–6; etc.).[3] In Figure 1b, recall rates of the critical list are depicted as a function of CONDITION (restudy, retrieval practice, distractor, semantic generation) and SERIAL POSITION (1-3, 4-6, 7-9, 10-12, 13-15, 16-18, 19-21, 22-24). A 4×8 ANOVA with the two factors revealed a main effect of CONDITION, $F(3,284) = 14.700$, $MSE = .282$, $p < .001$, partial $\eta^2 = .134$, which is consistent with the list-level analysis above, as well as a main effect of SERIAL POSITION, $F(7,1988) = 10.900$, $MSE = .089$, $p < .001$, partial $\eta^2 = .037$, reflecting that recall performance generally declined from the

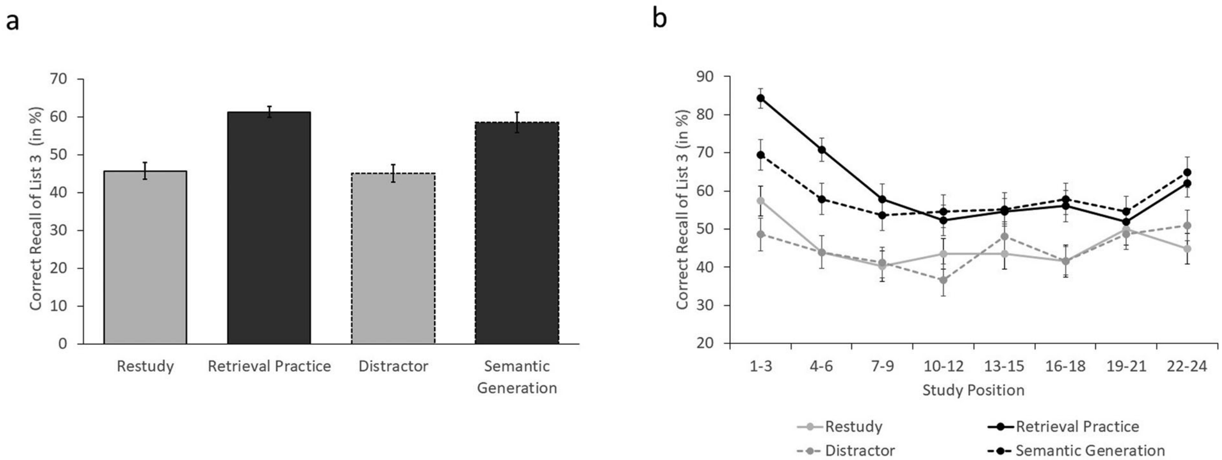

Figure 1. Recall rates of the critical list 3 in Experiment 1. a) On the list level, recall rates are shown for each of the four conditions (restudy, retrieval practice, distractor, semantic generation). (b) On the item level, recall rates are plotted as a function of CONDITION and SERIAL POSITION (1-3, 4-6, 7-9, 10-12, 13-15, 16-18, 19-21, 22-24). Error bars reflect standard errors.

earlier to the later bins. Critically, there was also a significant interaction between factors, $F(21,1988) = 2.101$, $MSE = .089$, $p = .002$, partial $\eta^2 = .022$. To examine this interaction in more detail, two separate 2 × 8 ANOVAS were conducted, in each of which we included the same eight bins (i.e., 1-3, 4-6, 7-9, 10-12, 13-15, 16-18, 19-21, 22-24) as in the above analysis, but either included (i) restudy and retrieval practice, or (ii) distractor and semantic generation.

First, when including the restudy and retrieval-practice conditions only, ANOVA showed significant main effects of CONDITION, $F(1,142) = 35.148$, $MSE = .199$, $p < .001$, partial $\eta^2 = .198$, and SERIAL POSITION, $F(7,994) = 9.269$, $MSE = .092$, $p < .001$, partial $\eta^2 = .061$, reflecting overall lower recall in the restudy than retrieval-practice condition and an overall decline in recall rates from the earlier to the later bins. Critically, there was also a significant interaction between factors, $F(7,994) = 2.889$, $MSE = .092$, $p = .005$, partial $\eta^2 = .020$, suggesting a more pronounced FTE for early than middle and late critical items. Indeed, when bins 1 and 2 (i.e., serial positions 1-6) were removed from the analysis, the interaction between factors was no longer evident, $F(5,710) = 1.357$, $MSE = .094$, $p = .238$, partial $\eta^2 = .009$, $B_{01} > 30$, thus supporting the view that recall performance of the early critical items was crucial for the interaction effect.

Second, when including the distractor and semantic-generation conditions only, ANOVA showed significant main effects of CONDITION, $F(1,994) = 14.465$, $MSE = .365$, $p < .001$, partial $\eta^2 = .092$, and SERIAL POSITION, $F(7,994) = 3.658$, $MSE = .085$, $p = .001$, partial $\eta^2 = .025$, reflecting overall higher recall in the semantic-generation condition than the distractor condition, and an overall decline in recall rates from the earlier to the later bins. There was, however, no significant interaction between factors, $F(7,994) = 1.100$, $MSE = .085$, $p = .360$, partial $\eta^2 = .008$, $B_{01} > 30$, suggesting that the magnitude of the semantic-generation induced FTE did not vary reliably with serial position. [4]

Table 1. Mean number of list-3 intrusions for Experiments 1 and 2 (standard errors in parenthesis).

Condition	Restudy	Retrieval Practice	Distractor	Semantic Generation
Experiment 1	0.33 (0.09)	0.15 (0.05)	0.78 (0.13)	0.62 (0.12)
Experiment 2	1.32 (0.17)	0.68 (0.10)	1.50 (0.30)	1.61 (0.21)

Discussion

The results of Experiment 1 replicate prior work by showing that, for unrelated study lists, a general (list-level) FTE can arise both when interim retrieval practice and when interim semantic generation are applied (Divis & Benjamin, 2014; Kliegl & Bäuml, 2021; Pastötter et al., 2011) and by showing a differential FTE, i.e., a greater FTE for the critical list's early items than its middle and late items (Pastötter et al., 2018). However, the differential FTE arose in response to retrieval practice but was absent in response to semantic generation, which induced an FTE that did not vary in size with the items' serial position in the list. The findings of Experiment 1 thus align with the view that retrieval practice, but not semantic generation, can induce an encoding reset.

Intrusions during recall of the critical list were numerically lower in the retrieval-practice and semantic-generation conditions, relative to their control conditions, i.e., the restudy and distractor conditions. However, those reductions were not statistically significant, which deviates from some prior studies (e.g., Divis & Benjamin, 2014; Szpunar et al., 2008) but aligns with our recent study, which applied the same study material (Kliegl & Bäuml, 2021). The absence of significant differences in the present study and our earlier study may be due to the fact that the number of intrusions was already quite low in the restudy and distractor (control) conditions, so that there was not much room for a reliable reduction in the number of intrusions in response to retrieval activity.

Experiment 2

The goal of Experiment 2 was to examine whether the result of a differential FTE in response to retrieval practice generalises from unrelated lists to categorised lists. If an encoding reset mediated the differential FTE in Experiment 1, the effect should not be restricted to unrelated lists. Indeed, if an encoding reset arises because retrieval practice exhausts the opportunities to review studied material, it should arise for different types of study material. Analogously, the absence of a differential FTE in response to semantic generation as found in Experiment 1 with unrelated lists should generalise to categorised lists. With regard to semantic generation, however, there should also not be a general (list-level) FTE, given that the semantic-generation induced general FTE has been found with unrelated, but not with categorised study lists (Kliegl & Bäuml, 2021).[5]

Methods

Participants

On the basis of the estimate reported in Experiment 1, 288 students at Regensburg University were recruited for Experiment 2 (mean age = 21.9 years) with 72 subjects in each of the four experimental conditions.

Material, design, and procedure

All experimental details were identical to Experiment 1, with the exception of study material: For Experiment 2, three interrelated lists with 24 words each were constructed (set A, set B, set C). Each list contained four German exemplars from six categories (Van Overschelde et al., 2004). The six categories were building parts, kitchen utensils, body parts, musical instruments, weather phenomenons, and types of fabric. Items' average taxonomic frequencies neither differed between categories, $F(5,66) < 1$, nor between lists, $F(2,69) < 1$. List order was counterbalanced across subjects. Presentation order of the items was random within lists. The study material was identical to the material applied in one of our earlier studies (Kliegl & Bäuml, 2021, Experiments 1b, 2b, 3b).

Results

List-level analysis of critical-list recall

Correct recall. On the list level, the percentage of correctly recalled critical items is shown for each of the four experimental conditions in Figure 2a. An overall ANOVA of the four conditions (restudy, retrieval practice, distractor, semantic generation) showed a significant effect of CONDITION, $F(3,284) = 17.705$, $MSE = .028$, $p < .001$, partial $\eta^2 = .158$. Pairwise comparisons revealed that while the difference of 15.5% between the restudy and retrieval practice condition was reliable, $F(1,142) = 51.272$, $MSE = .017$, $p < .001$, partial $\eta^2 = .265$, the difference of 3.8% between the distractor and semantic-generation condition was not, $F(1,142) = 1.298$, $MSE = .039$, $p = .256$, partial $\eta^2 = .009$, $B_{01} = 6.231$, thus demonstrating that only retrieval practice, but not semantic generation, induced a general FTE.

Intrusions. An overall ANOVA of the four conditions (restudy, retrieval practice, distractor, semantic generation) showed a significant effect of CONDITION on number of intrusions, $F(3,284) = 3.990$, $MSE = 3.121$, $p = .008$, partial $\eta^2 = .040$. Pairwise comparisons revealed that the difference of 0.64 intrusions between the restudy and retrieval-practice condition was significant, $F(1,142) = 10.263$, $MSE = 1.432$, $p = .002$, partial $\eta^2 = .067$, but the difference of 0.11 intrusions between the distractor and semantic-generation condition was not, $F(1,144) < 1$, $B_{01} = 11.452$ (see Table 1).

Item-level analysis of critical-list recall

In Figure 2b, recall rates of the critical list are depicted as a function of CONDITION (restudy, retrieval practice, distractor, semantic generation) and SERIAL POSITION (1-3, 4-6, 7-9, 10-12, 13-15, 16-18, 19-21, 22-24). A 4 × 8 ANOVA with the two factors revealed a main effect of CONDITION, $F(3,284) = 17.495$, $MSE = .224$, $p < .001$, partial $\eta^2 = .156$, which is consistent with the list-level analysis above, as well as a main effect of SERIAL POSITION, $F(7,1988) = 9.143$, $MSE = .087$, $p < .001$, partial $\eta^2 = .031$, reflecting that recall performance generally declined from the earlier to the later bins. Critically, there was a significant interaction between factors, $F(21,1988) = 2.097$, $MSE = .087$, $p = .003$, partial $\eta^2 = .022$. Like in Experiment 1, two separate 2 × 8 ANOVAS were conducted to examine the interaction in more detail, comparing (i) restudy and retrieval practice, and (ii) distractor and semantic generation.

First, when including the restudy and retrieval-practice conditions only, ANOVA showed significant main effects of CONDITION, $F(1,142) = 50.382$, $MSE = .134$, $p < .001$, partial $\eta^2 = .262$, and SERIAL POSITION, $F(7,994) = 8.418$, $MSE = .088$, $p < .001$, partial $\eta^2 = .056$, reflecting overall lower recall in the restudy than retrieval-practice condition and an overall decline in recall rates from the earlier to the later bins. Critically, there was also a significant interaction between factors, $F(7,994) = 2.666$, $MSE = .088$, $p = .010$, partial $\eta^2 = .018$, suggesting a more pronounced FTE for early than middle and late critical items. However, when bins 1 and 2 (i.e., serial positions 1-6) were removed from the analysis, the interaction between factors was no longer evident, $F(5,710) < 1$, $B_{01} > 30$, thus supporting

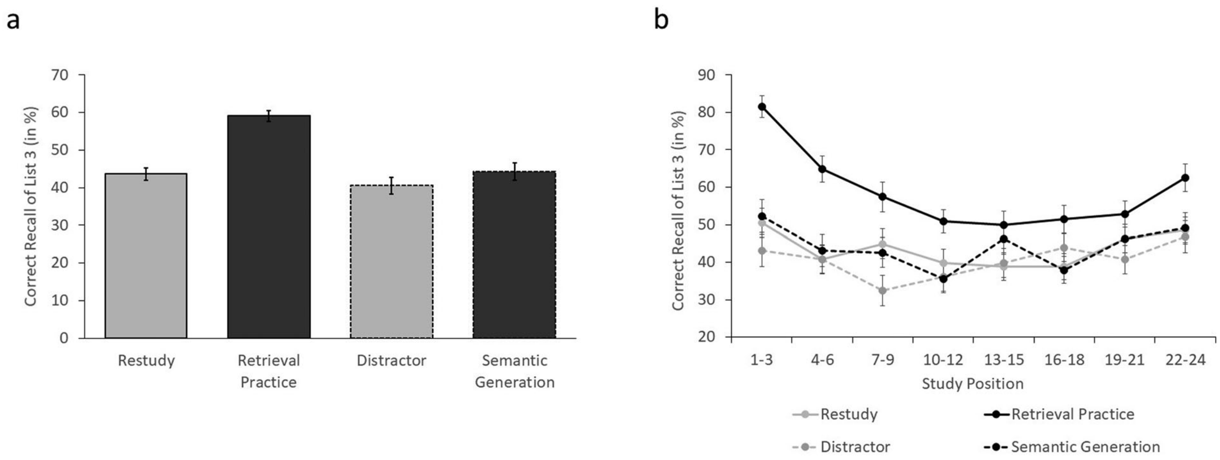

Figure 2. Recall rates of the critical list 3 in Experiment 2. a) On the list level, recall rates are shown for each of the four conditions (restudy, retrieval practice, distractor, semantic generation). (b) On the item level, recall rates are plotted as a function of CONDITION and SERIAL POSITION (1-3, 4-6, 7-9, 10-12, 13-15, 16-18, 19-21, 22-24). Error bars reflect standard errors.

the view that recall performance of the early critical items was crucial for the interaction effect.

Second, when including the distractor and semantic generation conditions only, there was a main effect of SERIAL POSITION, $F(7,994) = 3.085$, $MSE = .086$, $p = .003$, partial $\eta^2 = .021$, reflecting an overall decline in recall rates from the earlier to the later bins. There was, however, no main effect of CONDITION, $F(1,142) = 1.257$, $MSE = .314$, $p = .264$, partial $\eta^2 = .009$, $B_{01} = 6.361$, and no interaction between factors, $F(7,994) < 1$, $B_{01} > 30$. [6]

Discussion

The results of Experiment 2 generalise the findings of Experiment 1 by showing that retrieval practice induces both a general (list-level) FTE and a more pronounced FTE for early than middle and late critical items, thus indicating that the two effects do not depend on study material. These findings are consistent with the view that an encoding reset mediates the differential FTE and does so largely independent of study material. In contrast, semantic generation did not induce a general FTE and did also not induce a differential FTE. While the list-level result replicates prior work, which already indicated that semantic generation induces a general FTE with unrelated lists but not with categorised lists (Kliegl & Bäuml, 2021), the item-level result generalises the analogous finding of Experiment 1. Together, the results support the view that an encoding reset operates in response to retrieval practice but not in response to semantic generation.

Intrusions during recall of the critical list were significantly lower in the retrieval-practice than the restudy condition, but did not differ much between the semantic-generation and distractor conditions. Our recent study (Kliegl & Bäuml, 2021) showed similar effects of retrieval activities on number of intrusions when categorised lists were applied. However, this earlier study found no statistically significant difference between the retrieval practice restudy condition, probably due to study's lower statistical power.

General discussion

The results of the present experiments show a greater FTE for early than middle and late items of the critical list in response to retrieval practice, both when unrelated study material (Experiment 1) and when categorised study material (Experiment 2) were employed, indicating that type of study material is not a critical factor for the differential FTE. In contrast, no differential FTE was observed when interim semantic generation was applied, in both Experiment 1, in which the FTE was present on the list level, and Experiment 2, in which the FTE was absent on the list level. This pattern of results indicates that the differential (item-level) FTE is a retrieval-practice specific effect, whereas the general (list-level) FTE, at least with unrelated material, is not. The two FTEs can thus be dissociated experimentally.

The present results are consistent with the encoding-reset hypothesis, which holds that interim retrieval practice can abolish inattention and memory load and thus enable a more focused encoding of subsequently studied items. The critical assumption of this hypothesis is that the encoding reset arises in response to interim retrieval practice, because retrieval practice can leave subjects with the feeling that they are "finished with the list". In contrast, no such feeling should arise in response to interim semantic generation and thus no encoding reset should emerge. Our finding of a retrieval-specific differential FTE aligns with this assumption, although no direct measurement was provided of whether retrieval practice elicited a more pronounced feeling of closure or "being done" with the immediately preceding study list than semantic generation.

Prior work on the FTE suggested also roles of context change and strategy change for the FTE. Very recent work even alluded to the possibility that both mechanisms contribute to the FTE but play different roles in different experimental situations: whereas context change may be the critical mechanism underlying the general FTE with unrelated lists, strategy change may be the critical mechanism mediating the general FTE with categorised lists (Kliegl & Bäuml, 2021). In line with this two-factor account, it was shown that, with unrelated lists, the general FTE disappeared after both prolonged retention interval and prolonged lag, and was still present in response to interim semantic generation; these findings agree with the context-change explanation because context-change effects have been shown to be relatively short-lived but to arise in response to both retrieval practice and semantic generation (Divis & Benjamin, 2014; Pastötter et al., 2011). In contrast, with categorised study lists, the general FTE arose independent from the length of both retention interval and lag, and did not emerge in response to semantic generation; this finding aligns with the strategy-change explanation, if it is assumed that changes in encoding and/or retrieval strategy are lasting and interim semantic generation provides no information on the learning task that might promote the application of new and more effective encoding strategies (Chan, Manley, et al., 2018).

The present results support the two-factor account of the FTE by replicating the finding that, on the list level, a semantic-generation induced general FTE arises with unrelated study lists but not with categorised study lists. However, the present results go beyond the two-factor account by additionally demonstrating that, although both retrieval practice and semantic generation induce a general FTE with unrelated lists, retrieval practice – but not semantic generation – induces a differential FTE. This pattern is difficult to reconcile with the two-factor account because if context change mediated both the list-level and the item-level FTE with unrelated lists, semantic generation should not only induce the list-level FTE but also the item-level FTE, which was not the case. An encoding reset may thus provide an additional contribution to the FTE that goes beyond context change and strategy change. Consistent with this suggestion, Chan et al. (2020) found evidence that, with categorised lists, interim retrieval-practice can independently induce enhancements in both strategic *and* attentional processing.

An alternative explanation for the differential FTE is the output-order account, which holds that the differential FTE is the result of retrieval processes manifesting at test (Yang et al., 2021). Yang et al. found that while participants in the retrieval-practice condition tended to initiate their recall of the critical list with early list items, participants in the restudy condition tended to initiate their recall with late list items. Because first recalled items can impair recall of other items (Smith, 1971), recall of early list items in the retrieval-practice condition may have hindered recall of the later list items, and recall of late list items in the restudy condition may have impaired recall of the earlier list items, thus leading to a differential FTE. Such differences in output order, however, should not only enhance the FTE for early relative to middle and late items, but should also reduce the FTE for late relative to middle items, which is not what the results show. Moreover, Yang et al. (2021) found that the differential FTE survived when output order was controlled, suggesting that the differential FTE is not primarily attributable to differences in output order between practice conditions. [7]

Aside from encoding reset, context change and strategy change, further processes may contribute to the diverging effects of interim retrieval practice and interim semantic generation on recall of the last studied list. While, naturally, interim retrieval practice should lead to a comprehensive reactivation of the immediately preceding study list, interim semantic generation may in some cases also lead to reactivation of some items from the prior list (e.g., Hintzman, 2004), and the probability of such remindings could vary with type of study material. In the present study, such potential dynamics were not directly examined. Future research may address the issue and come up with predictions about whether and how such dynamics might affect recall performance of the last studied list.

Our findings suggest a parallel between the FTE and list-method directed forgetting. List-method directed forgetting refers to the observation that a cue to forget previously studied material can promote learning and memory of subsequently studied material. In a typical list-method directed forgetting task, participants study two item lists and, after study of list 1, they are either asked to forget the list (forget condition) or to keep remembering the list for a later test (remember condition). After study of list 2 – which is always to-be-remembered – participants are asked to recall the two lists' items irrespective of original cuing (Bjork, 1989). Forget-cued subjects typically show impaired recall of list 1 but increased recall of list 2. Moreover, analysis of items' serial position curves revealed that the forget cue induces a larger increase in recall for early list-2 items than for middle and late list-2 items (Pastötter et al., 2012; Pastötter & Bäuml, 2010). The presence of similar serial position effects in the list-method directed forgetting and FTE tasks may indicate that both a cue to forget previously studied material and retrieval practice of previously studied material induce an encoding reset. An encoding reset may thus constitute a more general phenomenon and be present over a wider range of experimental situations.

Conclusions

Prior work has demonstrated that both interim retrieval practice and interim semantic generation induce the general (list-level) FTE when unrelated study lists are applied, whereas retrieval practice only creates the effect with categorised study lists. This study replicates these findings as well as the finding that a differential FTE – with a larger FTE for early than middle and late items in the critical material – arises in response to retrieval practice when unrelated material is studied. The present results critically extend the prior work by showing that the differential FTE is present in response to retrieval practice but is absent in response to semantic generation. Moreover, this pattern held regardless of which material was studied, thus experimentally dissociating the general (list-level) from the differential (item-level) FTE. The findings are not easily captured by context-change or strategy-change explanations of the FTE and point to an additional role of an encoding reset for the FTE.

Open practices statement

The data and materials for all experiments are available at https://osf.io/duts3/, and none of the experiments was preregistered.

Notes

1. Certainly, retrieval practice will not always induce a feeling of closure. For instance, subjects may sometimes be experiencing a “tip-of-the-tongue” phenomenon, i.e., a feeling that retrieval is imminent (e.g., Brown & McNeill, 1966), which may motivate them to continue with their retrieval attempts.
2. The restudy condition is usually considered the appropriate control condition for the retrieval-practice condition, since both conditions involve repetition of immediately preceding study material; the distractor condition is usually considered the appropriate control condition for the semantic-generation condition, since neither condition involves repetition of previously studied material (e.g., Divis & Benjamin, 2014; Kliegl & Bäuml, 2021; Pastötter et al., 2011).
3. Critically, size of bins did not affect results, leading to the same conclusions regardless of whether data were analyzed with a six-bin selection (bin 1, items 1-4; bin 2, items 5-8; etc.) or a four-bin selection (bin 1, items 1-6; bin 2, items 7-12; etc.).
4. There was also a condition x serial position interaction when including retrieval practice and semantic generation conditions only, $F(7{,}994) = 2.269$, $MSE = .084$, $p = .027$, partial $\eta^2 = .016$ suggesting that retrieval practice induced a more pronounced primacy effect for critical items than semantic generation. Indeed, removing bins 1 and 2 from the analysis eliminated the interaction effect, $F(5{,}710) < 1$, $B_{01} > 30$.
5. A differential FTE may arise even in the absence of a general FTE since a differential FTE that transpires only for early, but not middle and late, items of the critical list in itself may not have sufficient leverage to induce a significant general FTE (see Pastötter & Bäuml, 2010).
6. Like in Experiment 1, there was also a CONDITION x SERIAL POSITION interaction when including retrieval practice and semantic generation conditions only, $F(7{,}994) = 2.744$, $MSE = .085$, $p = .008$, partial $\eta^2 = .019$, suggesting that retrieval practice induced a more pronounced primacy effect for critical items than semantic generation. Indeed, removing bins 1 and 2 from the analysis eliminated the interaction effect, $F(5{,}710) < 1.006$, $MSE = .084$, $p = .413$, partial $\eta^2 = .007$, $B_{01} > 30$.
7. Yang et al. (2021) also conducted a mediation analysis which revealed that the steepness of the serial position curve does not mediate the magnitude of the general FTE. While the researchers interpreted this result as evidence against the view that an encoding reset contributes to the general FTE at all, it is unclear whether the steepness of the serial position curve can be used as a pure indicator for the contribution of an encoding reset. Indeed, as noted above, there is evidence that further mechanisms may independently contribute to the general FTE, such as context change or strategy change, which may also affect the steepness of the curve – in particular by increasing recall rates of all study items and thus limiting the latitude for the curve’s steepness. We therefore argue that the results of this mediation analysis simply show that an encoding reset is not the sole mechanism contributing to the FTE.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This work was supported by Deutsche Forschungsgemeinschaft: [Grant Number BA 1382/18-1].

References

Aslan, A., & Bäuml, K.-H. T. (2016). Testing enhances subsequent learning in older but not in younger elementary school children. *Developmental Science*, *19*(6), 992–998. https://doi.org/10.1111/desc.12340

Bäuml, K.-H. T., & Kliegl, O. (2013). The critical role of retrieval processes in release from proactive interference. *Journal of Memory and Language*, *68*(1), 39–53. doi:10.1016/j.jml.2012.07.006

Bjork, R. A. (1989). Retrieval inhibition as an adaptive mechanism in human memory. In H. L. Roediger, III & F. I. M. Craik (Eds.), *Varieties of memory and consciousness: Essays in honour of endel tulving* (pp. 309–330). Erlbaum.

Brown, R., & McNeill, D. (1966). The “tip of the tongue” phenomenon. *Journal of Verbal Learning and Verbal Behavior*, *5*(4), 325–337. https://doi.org/10.1016/S0022-5371(66)80040-3

Chan, J. C., Manley, K. D., & Ahn, D. (2020). Does retrieval potentiate new learning when retrieval stops but new learning continues?. *Journal of Memory and Language*, *115*, 104150. https://doi.org/10.1016/j.jml.2020.104150

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. https://doi.org/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

Davis, S. D., & Chan, J. C. K. (2015). Studying on borrowed time: How does testing impair new learning?. *Journal of Experimental Psychology: Learning Memory and Cognition*, *41*(6), 1741–1754. https://doi.org/10.1037/xlm0000126

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. https://doi.org/10.3758/s13421-014-0425-y

Duyck, W., Desmet, T., Verbeke, L. P., & Brysbaert, M. (2004). Wordgen: A tool for word selection and nonword generation in Dutch, English, German, and French. *Behavior Research Methods, Instruments, & Computers*, *36*(3), 488–499. https://doi.org/10.3758/BF03195595

Estes, W. K. (1955). Statistical theory of spontaneous recovery and regression. *Psychological Review*, *62*(3), 145–154. https://doi.org/10.1037/h0048509

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. doi:10.3758/BRM.41.4.1149

Hintzman, D. L. (2004). Judgment of frequency versus recognition confidence: Repetition and recursive reminding. *Memory & Cognition*, *32*(2), 336–350. https://doi.org/10.3758/BF03196863

Karpicke, J. D. (2017). Retrieval-based learning: A decade of progress. In J. H. Byrne (Ed.), *Cognitive psychology of memory, Vol. 2 of learning and memory: A comprehensive reference* (pp. 1–26). Elsevier. http://doi.org/10.1016/B978-0-12-809324-5.21055-9

Karpicke, Jeffrey D., Butler, Andrew C., & Roediger III, Henry L. (2009). Metacognitive strategies in student learning: Do students practise retrieval when they study on their own?. *Memory*, *17*(*4*), 471–479. http://dx.doi.org/10.1080/09658210802647009

Kliegl, O., & Bäuml, K.-H. T. (2021). When retrieval practice promotes new learning – the critical role of study material. *Journal of Memory and Language*, *120*, 104253. https://doi.org/10.1016/j.jml.2021.104253

Kliegl, O., Pastötter, B., & Bäuml, K.-H. T. (2015). The contribution of encoding and retrieval processes to proactive interference. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *41*(6), 1778–1789. https://doi.org/10.1037/xlm0000096

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

Masson, M. E. (2011). A tutorial on a practical Bayesian alternative to null-hypothesis significance testing. *Behavior Research Methods*, *43*(3), 679–690. https://doi.org/10.3758/s13428-010-0049-5

Mensink, G. J., & Raaijmakers, J. G. (1988). A model for interference and forgetting. *Psychological Review*, *95*(4), 434. https://doi.org/10.1037/0033-295X.95.4.434

Pastötter, B., & Bäuml, K.-H. (2010). Amount of postcue encoding predicts amount of directed forgetting. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, *36*(1), 54–65. doi:10.1037/a0017406

Pastötter, B., & Bäuml, K.-H. T. (2014). Retrieval practice enhances new learning: The forward effect of testing. *Frontiers in Psychology*, *5*, 286. https://doi.org/10.3389/fpsyg.2014.00286

Pastötter, B., & Bäuml, K.-H. T. (2019). Testing enhances subsequent learning in older adults. *Psychology and Aging*, *34*(2), 242–250. https://doi.org/10.1037/pag0000307

Pastötter, B., Engel, M., & Frings, C. (2018). The forward effect of testing: Behavioral evidence for the reset-of-encoding hypothesis using serial position analysis. *Frontiers in Psychology*, *9*, 1197. https://doi.org/10.3389/fpsyg.2018.01197

Pastötter, B., Kliegl, O., & Bäuml, K.-H. T. (2012). List-method directed forgetting: The forget cue improves both encoding and retrieval of postcue information. *Memory & Cognition*, *40*(6), 861–873. https://doi.org/10.3758/s13421-012-0206-4

Pastötter, B., Weber, J., & Bäuml, K.-H. T. (2013). Using testing to improve learning after severe traumatic brain injury. *Neuropsychology*, *27*(2), 280–285. https://doi.org/10.1037/a0031797

Pastötter, Bernhard, Schicker, Sabine, Niedernhuber, Julia, & Bäuml, Karl-Heinz T. (2011). Retrieval during learning facilitates subsequent memory encoding. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *37*(*2*), 287–297. http://dx.doi.org/10.1037/a0021801

Pennebaker, J. W., & Chung, C. K. (2011). Expressive writing: Connections to physical and mental health. In H. S. Friedman (Ed.), *The Oxford handbook of health psychology* (pp. 417–437). Oxford University Press.

Raftery, A. E. (1995). Bayesian model selection in social research. *Sociological Methodology*, *25*, 111–163. doi:10.2307/271063

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

Roediger, H. L., III & Karpicke, J. D. (2006). Test-enhanced learning: Taking memory tests improves long-term retention. *Psychological Science*, *17*(3), 249–255. https://doi.org/10.1111/j.1467-9280.2006.01693.x

Roediger, H. L., III, Putnam, A. L., & Smith, M. A. (2011). Ten benefits of testing and their applications to educational practice. *Psychology of Learning and Motivation*, *55*, 1–36. https://doi.org/10.1016/B978-0-12-387691-1.00001-6

Roets, A., Van Hiel, A., & Cornelis, I. (2006). The dimensional structure of the need for cognitive closure scale: Relationships with "seizing" and "freezing" processes. *Social Cognition*, *24*(1), 22–45. https://doi.org/10.1521/soco.2006.24.1.22

Rohrer, Doug. (1996). On the relative and absolute strength of a memory trace. *Memory & Cognition*, *24*(*2*), 188–201. http://dx.doi.org/10.3758/BF03200880

Smith, A. D. (1971). Output interference and organized recall from long-term memory. *Journal of Verbal Learning and Verbal Behavior*, *10*(4), 400–408. https://doi.org/10.1016/S0022-5371(71)80039-7

Soderstrom, N. C., & Bjork, R. A. (2014). Testing facilitates the regulation of subsequent study time. *Journal of Memory and Language*, *73*, 99–115. https://doi.org/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 of the United States of America*, *110*(16), 6313–6317. https://doi.org/10.1073/pnas.1221764110

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

Unsworth, N., Brewer, G. A., & Spillers, G. J. (2013). Focusing the search: Proactive and retroactive interference and the dynamics of free recall. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *39*(6), 1742–1756. https://doi.org/10.1037/a0033743

Van Overschelde, J. P., Rawson, K. A., & Dunlosky, J. (2004). Category norms: An updated and expanded version of the norms. *Journal of Memory and Language*, *50*(5), 289–335. https://doi.org/10.1016/j.jml.2003.10.003

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. https://doi.org/10.3758/s13423-011-0085-x

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. https://doi.org/10.3758/s13423-011-0140-7

Wixted, J. T., & Rohrer, D. (1993). Proactive interference and the dynamics of free recall. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *19*(5), 1024. https://doi.org/10.1037/0278-7393.19.5.1024

Yang, C., Potts, R., & Shanks, D. R. (2017). The forward testing effect on self-regulated study time allocation and metamemory monitoring. *Journal of Experimental Psychology: Applied*, *23*(3), 263–277. https://doi.org/10.1037/xap0000122

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), 1–9. doi:10.1038/s41539-017-0017-2

Yang, C., Zhao, W., Luo, L., Sun, B., Potts, R., & Shanks, D. R. (2021). Testing potential mechanisms underlying test-potentiated new learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*. Advance online publication. https://doi.org/10.1037/xlm0001021