Journal of Memory and Language 120 (2021) 104253

Contents lists available at ScienceDirect

Journal of Memory and Language

journal homepage: www.elsevier.com/locate/jml

When retrieval practice promotes new learning – The critical role of study material

Oliver Kliegl [*], Karl-Heinz T. Bäuml

Department of Experimental Psychology Regensburg University, Germany

ARTICLE INFO

Keywords:
Recall
Retrieval practice
Semantic generation
Retention interval
Lag

ABSTRACT

The forward testing effect (FTE) refers to the finding that retrieval practice of previously studied information can facilitate learning and memory of newly studied information. The goal of the present set of six experiments was to examine whether the FTE is influenced by study material. We replicated prior work by showing that the FTE can arise with both unrelated and categorized item lists. Going beyond the prior work, we found that parallel FTEs for the two types of lists arose only for short retention interval and when the lag between study of the previous lists and study of the final critical list was also short. When there was a prolonged retention interval or a prolonged lag, the FTE was observed with categorized lists but disappeared with unrelated lists. Moreover, semantic generation of extra-list items interspersed between study of the single lists produced an FTE with unrelated lists but not with categorized lists. These findings on the critical role of study material for the FTE are consistent with a two-factor explanation of the FTE, which assumes contributions of both strategy change and context change for the FTE. The account suggests that the FTE is mainly driven by strategy change with categorized material and is mainly driven by context change with unrelated material.

When Retrieval Practice Promotes New Learning – The Critical Role of Study Material

Over a hundred years of memory research have substantiated the view that testing – or retrieval practice – can afford tremendous benefits for learning and long-term retention. Studies on the so-called testing effect in particular have demonstrated that active retrieval of previously learned material can improve later memory of the practiced material much more than repeated presentation of the same material. This finding has turned out to be very robust and has been observed across a wide range of experimental situations and subject groups (see Roediger & Butler, 2011; Rowland, 2014).

However, retrieval practice improves not only retention of the practiced material itself, but also fosters the learning and memory of subsequently encountered, new material. This effect was first demonstrated in a study by Szpunar, McDermott, and Roediger (2008), in which subjects were asked to learn five lists of words successively in anticipation of a final cumulative recall test and, immediately after the presentation of lists 1 through 4, were asked to solve simple mathematical problems (distractor condition), study the word lists once again (restudy condition), or try to recall the words of the immediately preceding list (retrieval-practice condition). After learning list 5, all subjects were asked to recall the words of this critical final list. Results showed that subjects in the retrieval-practice condition remembered more words from list 5 and showed less intrusions of words from lists 1 through 4 than subjects in the other two conditions. This effect is often referred to as the forward testing effect (FTE) in contrast to the classic testing effect, which has sometimes been referred to as the backward testing effect (see Pastötter & Bäuml, 2014). Like the classic testing effect, the FTE is a very general effect. It has been found in both lab-based studies and educational settings, and has been shown to arise across a variety of study materials, including word lists, paired associates (Weinstein, McDermott, & Szpunar, 2011), prose material (Wissman, Rawson, & Pyc, 2011), and videos (Szpunar, Khan, & Schacter, 2013). The FTE has also 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, Weber, & Bäuml, 2013).

All item materials that were applied in the present experiments as well as all data are available on the Open ScienceFramework (https://osf.io/73va8/).

[*] Corresponding author. Department of Experimental Psychology, Regensburg University, 93040 Regensburg, Germany.
E-mail address: oliver.kliegl@ur.de (O. Kliegl).

https://doi.org/10.1016/j.jml.2021.104253
Received 17 November 2020; Received in revised form 14 April 2021; Accepted 16 April 2021

Explanations of the FTE

Explanations of the FTE include context-change explanations and strategy-change explanations (for an overview, see Chan, Meissner et al., 2018; for multi-factor accounts, see also General Discussion below). Context-change explanations are based on the widespread view that when we encode information we also store details about the mental context in which this information is encountered (Estes, 1955; Mensink & Raaijmakers, 1988). Critically, retrieval activities between the study of lists are assumed to drive mental context change, thereby contextually segregating newly from previously learned material (Jang & Huber, 2008; Shiffrin, 1970). This segregation should reduce interference from the previously learned material at the time of test, and thus enable more focused memory search for the newly learned material (Bäuml & Kliegl, 2013; Pastötter, Schicker, Niedernhuber, & Bäuml, 2011; Szpunar et al., 2008). In contrast, strategy-change explanations are based on the idea that retrieval practice can induce subjects to employ new, and potentially more effective strategies for further learning. Retrieval practice can indeed provide important information for the further tests, such as information about the test format or the presence of retrieval cues, on the basis of which encoding and retrieval strategies may be optimized (Chan, Manley et al., 2018; Davis & Chan, 2015; Soderstrom & Bjork, 2014).

Until recently, research on the FTE has largely considered the two explanations in isolation and has used them to account for findings from single experiments. However, in two recent studies (Chan, Manley et al., 2018; Chan, Meissner et al., 2018), Chan and colleagues provided a more comprehensive explanation of the FTE. This explanation is based on the assumption that the FTE is produced by a change in encoding and retrieval strategies without any major contribution of context change (for a similar proposal, see Cho, Neely, Crocco, & Vitrano, 2017). Partly this strategy-change explanation is based on findings from a recent series of experiments on the FTE (Chan, Manley et al., 2018), and partly it is based on findings from a meta-analysis on retrieval-induced new learning, which included results on both the FTE and related experimental situations (Chan, Meissner et al., 2018).

In their experimental series, Chan, Manley et al. (2018) asked subjects to study four categorized word lists, each list consisting of three exemplars from five semantic categories. Across lists, the same five categories were used. Between the consecutive lists, subjects were asked to either engage in unrelated distractor activities, restudy the immediately preceding list, or retrieve the list they had just learned. Chan, Manley et al. found the typical FTE with better memory after retrieval practice compared to the other two conditions. More important, the researchers showed that the FTE was largely preserved when the retention interval after learning of list 4 was extended from 1 min to 25 min, and when the lag between study of list 4 and study of the prior lists 1–3 was increased from 1 min to 25 min. These findings are consistent with the strategy-change view because the effects of a strategy shift should still benefit the encoding and retrieval of the critical (final) list when lag or retention interval were increased, at least when adjustments in strategy are assumed to be lasting. Chan, Manley et al. provided further evidence in favor of the strategy-change explanation by showing that the FTE was accompanied by an improved semantic organization of the final target list (list 4), as subjects in the retrieval-practice condition showed a stronger propensity to cluster their recall on the basis of the items' category membership.

The findings, however, challenge the context-change explanation. This explanation predicts that the FTE should be largely eliminated after prolonged retention interval as well as after lagged learning. Prolonged retention interval typically eliminates any effects of context change (Abel & Bäuml, 2017, 2019; Divis & Benjamin, 2014), indicating that there should barely be any beneficial effects of retrieval-induced context change on list-4 recall after prolonged retention interval. Furthermore, lagged learning should induce context change also in the absence of retrieval practice (Estes, 1955;Mensink & Raaijmakers, 1988), suggesting that the retrieval-induced context change should not enhance list-4 recall much further after prolonged lag.

On the basis of their findings, Chan, Manley et al. (2018) suggested that strategy change is *the* critical mechanism underlying the FTE, a conclusion also supported by meta-analytical findings on retrieval-induced new learning (see Chan, Meissner et al., 2018). A direct consequence of this theoretical view is also that the magnitude of the FTE should vary with study material and the FTE be, for instance, more pronounced in categorized lists than in unrelated lists, in which the items are largely unconnected, both within lists and across lists. Indeed, since in unrelated lists, the items' categories are not repeated across lists, this material should make it more difficult to improve semantic organization and thus prevent the FTE from unfolding. Again, this view is supported by meta-analytical findings, which indicate that the FTE is somewhat smaller with unrelated than categorized lists (see Chan, Meissner et al., 2018).

A two-factor acccount of the FTE

However, not all findings from the literature on the FTE align with the strategy-change explanation. For instance, studies employing unrelated lists as study material found that the FTE is not limited to situations in which the previously learned material is tested, but can also arise in response to a semantic generation task, in which subjects, between the study of the single lists, are asked to produce as many exemplars as possible from extra-list categories (e.g., PROFESSIONS; Divis & Benjamin, 2014; Pastötter et al., 2011). This observation is not covered by the strategy-change explanation, which suggests that the FTE is bound to retrieval practice; in fact, whereas retrieval practice can provide information on forthcoming tests, semantic generation does not. The observation, however, is consistent with the context-change account, since semantic generation has been found to induce mental context change in several experimental tasks (e.g., Jang & Huber, 2008; Rupprecht & Bäuml, 2017; but see Weinstein, 2015).[1]

Another finding from Divis and Benjamin's (2014) study challenges the strategy-change account. Employing unrelated word lists, this study indicated that an FTE, which is created by semantic generation, can disappear with longer retention interval. This result contrasts with Chan, Manley et al.'s (2018) result that the FTE persists with categorized lists. Moreover, it is consistent with the context-change account, according to which the FTE should be transient. Finally, some findings on the FTE show that the size of the FTE with unrelated lists can be quite similar to the size of the FTE with categorized lists (e.g., Aslan & Bäuml, 2016; Pastötter et al., 2011; Szpunar et al., 2008), which is difficult to reconcile with strategy change, according to which the FTE should be more pronounced in categorized than unrelated lists. All of these findings indicate that the results reported by Chan and colleagues with categorized lists may not generalize to unrelated lists. In particular, they suggest that both strategy change and context change may contribute to the FTE, but with different relative contributions of the two cognitive mechanisms for different study material. Such a two-factor account of

[1] Weinstein (2015) reported a failure to find evidence that interpolated autobiographical and interpolated semantic generation induce an FTE, suggesting that autobiographical and semantic generation may not trigger context change. These findings contrast with the results of a number of studies, showing that imagination tasks – in which subjects are asked to generate autobiographical memories like their parents' home or their last vacation – as well as semantic-generation tasks can be highly effective at inducing context change (e.g., Divis & Benjamin, 2014; Jang & Huber, 2008; Pastötter & Bäuml, 2007; Pastötter et al., 2011; Rupprecht & Bäuml, 2017; Sahakyan & Kelley, 2002). In particular, Divis and Benjamin (2014) demonstrated that, when unrelated study lists are employed, semantic generation induces an FTE for the final list as well as forgetting of the first list items, which is exactly the pattern one would expect if semantic generation induced context change. Similar results were reported for autobiographical generation (e.g., Sahakyan & Kelley, 2002).

the FTE is suggested next.

The account is based on two proposals. The one proposal is that retrieval practice induces mental context change. Such context change should be effective with unrelated lists. It should lead to enhanced segregation of the study lists, thus enable a more focused memory search for the critical list and improve recall of the list. In contrast, with categorized lists as study material, such context change may not improve recall performance. The reason is that the repetition of the categories across study lists may reinstate the study context of the previous list(s) during learning of each single list (Jonker, Seli, & MacLeod, 2013; Wirth & Bäuml, 2020) and thus keep context relatively stable across lists. As a result, induced context change should contribute to the FTE mainly with unrelated lists and much less with categorized lists. The second proposal is that retrieval practice does not only induce mental context change but can also optimize encoding and retrieval strategies. Such optimization should arise mainly with categorized lists. With such lists, retrieval practice provides performance-relevant information – for instance, about the structure of the study material – and may thus shift subjects' strategy by encoding and retrieving the items of the subsequent lists on the basis of their category affiliation. A comparable shift in strategy should not occur with unrelated lists, which show low pre-existing semantic relationships between the study lists. As a result, strategy change should contribute to the FTE mainly with categorized lists and much less with unrelated lists. Following the two proposals, the FTE should therefore arise with both categorized and unrelated lists, but the cognitive mechanisms driving the effect should vary with study material.

On the basis of this two-factor account, qualitatively different FTEs should arise for different study material. Two expectations stand out. The one expectation relates to the question of whether not only retrieval practice but also semantic generation can induce an FTE. Because, with unrelated lists, the FTE should be due to context change mainly, both retrieval practice and semantic generation should induce an FTE with unrelated lists. In contrast, because mainly strategy change should mediate the FTE with categorized lists, retrieval practice, but not semantic generation, should induce an FTE with this type of list. The other expectation relates to the possible role of retention interval and the possible role of lag between study of the initial list(s) and study of the critical list for the FTE. If primarily context change mediated the FTE with unrelated lists, a reduced, or even eliminated, FTE should be expected for this material after both prolonged retention interval and prolonged lag. In contrast, the FTE should be largely maintained across retention intervals and lags with categorized lists, if mainly strategy change mediated the effect for this material.

The present study

The present study addresses the role of study material for the FTE by examining separately for categorized and unrelated item lists whether (i) the FTE is specific to retrieval practice or generalizes to semantic generation, (ii) the FTE is maintained in size when the retention interval between study and test of the final list is increased, and (iii) the FTE is maintained in size when the lag between study of the initial lists and study of the final list is increased. For categorized lists, the two-factor account predicts that the FTE is specific to retrieval practice (and thus does not generalize to semantic generation), and is still present when retention interval and lag are increased. In contrast, for unrelated lists, the account predicts that the FTE arises after both retrieval practice and semantic generation and largely disappears when retention interval or lag are increased.

The results of six experiments are reported, in each of which we applied the three-list version of the typical FTE task (Bäuml & Kliegl, 2013;Pastötter, Engel, & Frings, 2018). In Experiments 1a and 1b, we examined whether the FTE is specific to retrieval practice or generalizes to semantic generation. While participants in the retrieval-practice condition were asked to recall as many list-1 and list-2 items as possible immediately after studying the respective lists, participants in the semantic-generation condition were prompted to name as many exemplars as possible from extra-list categories between study of the single lists. Experiment 1a employed unrelated lists and Experiment 1b categorized lists. Experiments 2a and 2b compared the magnitude of the FTE across a short 1-min and a longer 25-min retention interval between study and test of the critical list. During both retention intervals, participants engaged in unrelated distractor activities. Experiment 2a employed unrelated lists and Experiment 2b categorized lists. Finally, in Experiments 3a and 3b, we compared the size of the FTE across a short 1-min and a longer 25-min lag between study of the initial lists and study of the critical list. Participants again engaged in unrelated distractor activities during both lag intervals. Experiment 3a employed unrelated lists and Experiment 3b categorized lists.

Experiments 1a and 1b

All procedural details of Experiments 1a and 1b were identical with the only exception of the material that was presented at study. For this reason, the methods for the two experiments are reported together below. The goal of Experiments 1a and 1b was to examine whether the FTE is specific to retrieval practice or arises in response to semantic generation as well. To this end, we asked subjects to study three item lists. After study of lists 1 and 2, subjects were either asked to solve simple arithmetic problems (distractor condition), study the immediately preceding list once again (restudy condition), recall the words of the immediately preceding list (retrieval-practice 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. For this test, both recall totals and intrusions were measured.

Experiment 1a employed unrelated lists. On the basis of the two-factor account, an 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). Indeed, both types of retrieval activity should induce context change, thus enhance segregation between study lists and facilitate recall of the final list at the time of test. Experiment 1b employed categorized lists. Here, the two-factor account suggests that retrieval practice, but not semantic generation, should induce an FTE. Indeed, retrieval practice, but not semantic generation, should lead to an optimization of encoding and retrieval strategies for the subsequently learned (target) material and induce an FTE.

Method

Participants. Sample size in Experiments 1a and 1b – as well as in the remaining experiments reported in this manuscript – was determined on the basis of a power analysis that followed the settings of the power analysis reported in Chan, Manley et al. (2018). In particular, using the meta-analytic effect size of the FTE ($g = 0.75$, Chan, Meissner et al., 2018) and setting α to.05 and β to.15, 34 participants were recommended for each between-subjects condition. On the basis of this estimate, 144 students at Regensburg University were recruited for both Experiment 1a (mean age = 21.9 years) and Experiment 1b (mean age = 22.8 years), with 36 subjects in each of the four experimental conditions of the two experiments. 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 all six experiments were tested in person. Data collection was finished before the COVID-19 outbreak.

Material. For Experiment 1a, 72 unrelated German nouns of medium frequency were drawn from the CELEX database (Duyck, Desmet, Verbeke, & Brysbaert, 2004). For each participant, items were assigned randomly to three lists consisting of 24 items each. For Experiment 1b, 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, Rawson, & Dunlosky, 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.

Design. Experiments 1a and 1b both had a 2 × 2 design with the between-subjects factors of PRACTICE (present vs. absent) and RETRIEVAL (present vs. absent). Practice was present but retrieval was absent in the restudy condition, in which immediately after study of lists 1 and 2, participants were asked to study the immediately preceding list once more; both practice and retrieval were present in the retrieval-practice condition, in which immediately after study of lists 1 and 2, participants were asked to recall as many items as possible from the preceding list; both practice and retrieval were absent in the distractor condition, in which immediately after study of lists 1 and 2, subjects solved simple arithmetic tasks; finally, practice was absent but retrieval was present in the semantic-generation condition, in which participants were asked immediately after study of lists 1 and 2 to produce as many exemplars as possible from a given category.

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, they would ultimately be tested on all study lists.

At the start of the experiment, the items of the three lists were visually presented at the center 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 list 3 and the backward-counting task, participants in all conditions were asked to write down as many items as possible from list 3 on a piece of paper. They were given 120 s for this free-recall task. Following list-3 recall, 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 of Experiment 1a

For all experiments, we provide Bayes factors (B_{01}) – which reflect the odds in favor of the null hypothesis over the alternative hypotheses – when a finding does not reach conventional level of statistical significance (i.e., α =.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, a B_{01} ranging from 10–30 as strong evidence for the null hypothesis, and a B_{01} ranging from 30–100 as very strong evidence for the null hypothesis (Masson, 2011; Raftery, 1995).

List-3 recall

Correct recall. Fig. 1a (left panel) shows the percentage of correctly recalled list-3 items for each of the four experimental conditions. A 2 x 2 ANOVA with the between-subjects factors of PRACTICE (present vs. absent) and RETRIEVAL (present vs. absent) revealed no significant main effect of PRACTICE, $F(1,140) < 1$, $B_{01} = 8.155$, but a main effect of RETRIEVAL, $F(1,140) = 21.262, MSE = .033, p < .001$, partial $\eta^2 = .132$, reflecting better retention in the presence than absence of retrieval (65.6% vs. 51.6%). Critically, there was no significant interaction between the two factors, $F(1, 140) < 1$, $B_{01} = 11.634$, with the Bayes factor indicating that the data were almost twelve times more probable under the null hypothesis than the alternative hypothesis. Thus, both retrieval practice and semantic generation induced an FTE.

Intrusions. All items that participants produced during the list-3 recall test that did not belong to list 3 were counted as intrusions, i.e., words recalled from lists 1 or 2, or words that were not part of any of the study lists. A 2 x 2 ANOVA with the factors of PRACTICE and RETRIEVAL revealed no main effects of PRACTICE, $F(1,140) < 1$, $B_{01} = 9.479$, or RETRIEVAL, $F(1,140) = 2.711, MSE = .740, p = .102$, partial $\eta^2 = .019$, $B_{01} = 3.016$, and no significant interaction between the two factors, $F(1, 140) < 1$, $B_{01} = 11.942$ (see Table 1).

Recall across lists

For participants in the retrieval-practice condition, we also examined recall performance across the three lists. A repeated measures ANOVA with the within-subjects factor of LIST (list 1, list 2, list 3) showed no significant main effect, $F(2,70) = 2.196, MSE = .010, p = .119$, partial $\eta^2 = .059$, $B_{01} = 8.095$, reflecting that mean recall remained relatively stable across lists (list 1 = 63.4%, list 2 = 62.6%, list 3 = 67.2 %).

Results of Experiment 1b

List-3 recall

Correct recall. Fig. 1b shows the percentage of correctly recalled list-3 items for each of the four experimental conditions. A 2 x 2 ANOVA with the between-subjects factors of PRACTICE (present vs. absent) and RETRIEVAL (present vs. absent) revealed main effects of PRACTICE, $F(1,140) = 5.764$, $MSE = .039, p = .018$, partial $\eta^2 = .040$, and RETRIEVAL, $F(1, 140) = 8.177, MSE = .039, p = .005$, partial $\eta^2 = .055$, reflecting better recall in the presence than the absence of practice (53.1% vs. 45.2%) and better recall in the presence than the absence of retrieval (53.8% vs. 44.4%). Critically, there was also a significant interaction between the two factors, $F(1, 140) = 5.430$, $MSE = .039$, $p = .021$, partial $\eta^2 = .037$. Consistently, planned comparisons between the retrieval-practice and restudy conditions showed a reliable difference in recall rates (61.6% vs. 44.6%), $t(70) = 4.017, p < .001, d = 0.947$, whereas no such difference arose between the semantic-generation and distractor conditions (46.1% vs. 44.3%), $t(70) < 1$, $B_{01} = 8.016$, thus suggesting that retrieval induced an FTE only when the retrieval activity involved practice of the previously studied list. There was no reliable difference in

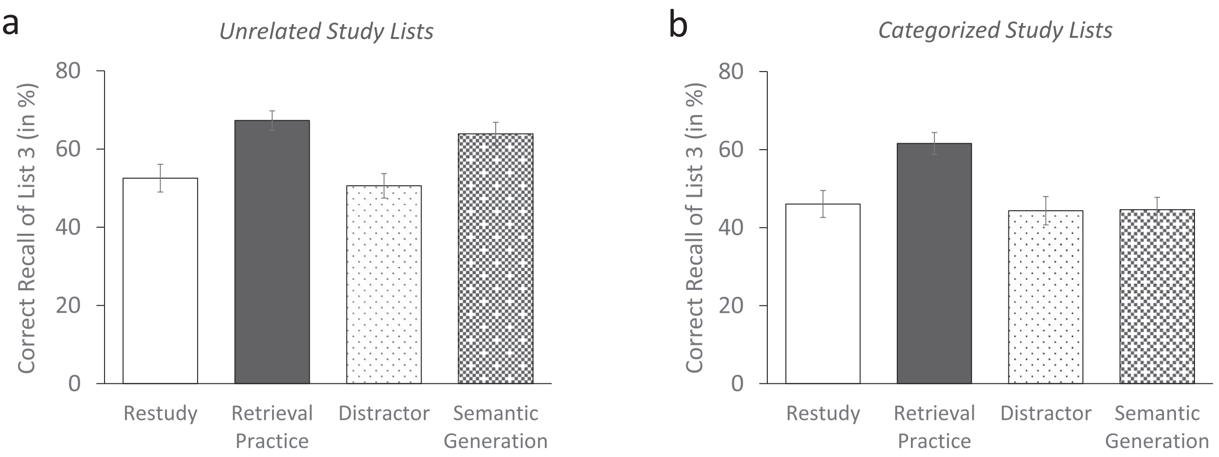

Fig. 1. (a) Results of Experiments 1a. Mean correct list-3 recall as a function of condition (restudy, retrieval practice, distractor, semantic generation) for lists of unrelated items. (b) Results of Experiments 1b. Mean correct list-3 recall as a function of condition (restudy, retrieval practice, distractor, semantic generation) for lists of categorized items. Error bars represent standard errors.

Table 1
Mean number of list-3 intrusions and mean adjusted-ratio-of-clustering scores of list 3 for Experiments 1–3 (standard errors in parenthesis). RI = retention interval.

	Intrusions				*Adjusted Ratio of Clustering*			
Condition	*Restudy*	*Retrieval Practice*	*Distractor*	*Semantic Generation*	*Restudy*	*Retrieval Practice*	*Distractor*	*Semantic Generation*
Experiment 1a	0.56 (0.18)	0.33 (0.08)	0.67 (0.17)	0.42 (0.12)	–	–	–	–
Experiment 1b	1.72 (0.27)	1.00 (0.21)	1.83 (0.28)	1.64 (0.25)	0.43 (0.06)	0.70 (0.05)	0.34 (0.06)	0.32 (0.07)
Condition	*Restudy (1-min RI)*	*Retrieval Practice (1-min RI)*	*Restudy (25-min RI)*	*Retrieval Practice (25-min RI)*	*Restudy (1-min RI)*	*Retrieval Practice (1-min RI)*	*Restudy (25-min RI)*	*Retrieval Practice (25-min RI)*
Experiment 2a	0.94 (0.17)	0.50 (0.13)	2.19 (0.64)	1.17 (0.22)	–	–	–	–
Experiment 2b	1.69 (0.31)	1.03 (0.18)	4.03 (0.71)	1.78 (0.21)	0.33 (0.08)	0.58 (0.07)	0.34 (0.07)	0.60 (0.06)
Condition	*Restudy (1-min Lag)*	*Retrieval Practice (1-min Lag)*	*Restudy (25-min Lag)*	*Retrieval Practice (25-min Lag)*	*Restudy (1-min Lag)*	*Retrieval Practice (1-min Lag)*	*Restudy (25-min Lag)*	*Retrieval Practice (25-min Lag)*
Experiment 3a	0.72 (0.19)	0.50 (0.15)	0.31 (0.10)	0.50 (0.13)	–	–	–	–
Experiment 3b	1.06 (0.16)	0.81 (0.15)	0.69 (0.13)	0.92 (0.18)	0.48 (0.07)	0.61 (0.06)	0.40 (0.08)	0.59 (0.06)

recall between the restudy and distractor conditions (44.6% vs. 44.3%), $t(70) < 1$, $B_{01} = 8.475$, whereas recall was significantly higher in the retrieval-practice than the semantic-generation condition (61.6% vs. 46.1%), $t(70) = 3.485, p = .001, d = 0.821$.

Intrusions. Regarding intrusions, a 2 x 2 ANOVA with the factors of PRACTICE and RETRIEVAL revealed no significant main effects of PRACTICE, $F(1, 140) = 3.262, MSE = 2.318, p = .073$, partial $\eta^2 = .023$, $B_{01} = 3.937$, and RETRIEVAL, $F(1, 140) = 2.184, MSE = 2.318, p = .142$, partial $\eta^2 = .015$, $B_{01} = 2.285$, and no significant interaction between the two factors, $F(1,140) = 1.081, MSE = 2.318, p = .300$, partial $\eta^2 = .008$, $B_{01} = 9.445$ (see Table 1).

Clustering in recall. To examine how interpolated retrieval activities affected participants' use of strategies, we calculated the likelihood with which participants clustered related items together during recall using the adjusted-ratio-of-clustering (ARC, Roenker, Thompson, & Brown, 1971). ARC indicates the probability with which related items are recalled successively. ARC scores can take values between −1 and 1, with positive ARC scores reflecting above chance clustering, 0 indicating chance level clustering, and negative scores indicating below chance clustering. In this analysis, undefined ARC scores arise when only one item is recalled from each category or when all of the recalled items are from the same category. Such undefined ARC scores occurred only very infrequently in the present experiment (and also in our Experiments 2b and 3b) and were substituted with the value 0 (see also Chan, Manley et al., 2018).

Table 1 shows list-3 ARC scores for each of the four experimental conditions. A 2 x 2 ANOVA with the factors of PRACTICE and RETRIEVAL revealed main effects of PRACTICE, $F(1, 140) = 15.477$, $MSE = .128$, $p < .001$, partial $\eta^2 = .100$, and RETRIEVAL, $F(1, 140) = 3.985, MSE = .128, p = .048$, partial $\eta^2 = .028$, reflecting higher ARC scores in the presence than the absence of practice (.57 vs..33) and higher ARC scores in the presence than the absence of retrieval (.51 vs..39). Critically, there was also a significant interaction between the two factors, $F(1,140) = 5.834, MSE = .128, p = .017$, partial $\eta^2 = .040$. Indeed, while planned comparisons between the retrieval-practice and restudy conditions showed a reliable difference in ARC scores (0.70 vs. 0.43), $t(70) = 3.348, p = .001, d = 0.789$, no difference in scores arose between the semantic-generation and distractor conditions (0.32 vs. 0.34), $t(70) < 1$, $B_{01} = 8.124$.

Recall across lists

For participants in the retrieval-practice condition, we also examined recall performance and clustering (ARC) scores across lists. Recall across lists was analyzed using a repeated measures ANOVA with the within-subjects factor of LIST (list 1, list 2, list 3). The analysis revealed no significant effect, $F(2, 70) < 1$, $B_{01} = 38.987$, reflecting that mean recall remained relatively stable across lists (list 1 = 59.6%, list 2 = 62.5%, list 3 = 61.6%). ARC scores across lists were analyzed using the same repeated measures ANOVA, which showed that ARC scores rose across lists, $F(1,35) = 11.646, MSE = .158, p = .002$, partial $\eta^2 = .250$, with the ARC score increasing from.41 in list 1 to.61 in list 2 and.70 in list 3.

Discussion

Consistent with prior work, the results of the two experiments show retrieval-practice effects for both unrelated and categorized study material (e.g., Szpunar et al., 2008), with higher recall for the critical final list when there was retrieval practice on the previously studied lists than when these lists were restudied. More important, the results also showed a beneficial effect for the critical list when there was semantic generation after each of the previous lists, relative to a condition, in which subjects were engaged in neutral distractor activities. This effect, however, was present only when unrelated lists had been studied and was absent when categorized lists had been learned. Retrieval-practice specificity thus varied with material.

The results are consistent with the suggested two-factor account of the FTE. According to this account, both strategy change and context change can contribute to the FTE, but mainly strategy change should contribute to the effect with categorized lists and mainly context change contribute to the effect with unrelated lists. Because beneficial effects of strategy change should arise in response to retrieval practice but not in response to semantic generation, and beneficial effects of context change emerge after both retrieval activities, the account can explain the different effects of semantic generation for the two types of item lists. At the same time, the results of the two experiments challenge explanations of the FTE which rely solely on one of the two cognitive mechanisms.

The two-factor account assumes that – just like retrieval practice – semantic generation should "always" induce context change, but with categorized lists, the repetition of categories across lists may induce context reinstatement, thereby hindering segregation of the single study lists and preventing the FTE. Alternatively, the finding that semantic generation did not induce an FTE with categorized lists may also be explained by assuming that semantic generation can trigger recall of previously studied lists in participants who have learned categorized lists.[2] For instance, a participant might come up with *cook* when asked to generate exemplars from the category PROFESSIONS, which might remind her of exemplars from the previously studied category KITCHEN UTENSILS, such as *plate*, *mixer*, or *spoon*. If such reminding occurred frequently during semantic generation and occurred mostly for categorized and much less for unrelated study lists, then the induced context change might be reduced, or be even prevented altogether, with categorized lists. Importantly, however, both the original and this alternative explanation for why semantic generation did not induce an FTE with categorized lists are consistent with the suggested two-factor account, differing only in whether the putative reminding is supposed to occur during subsequent list encoding or already earlier in the course of the semantic generation.

The present two-factor account includes the assumption that semantic generation can induce context change. This assumption is based on prior work that demonstrated generation-induced context change in a number of experimental tasks, including the list-before-last paradigm (Jang & Huber, 2008), the retrieval-practice task (Rupprecht & Bäuml, 2017), and the FTE task (Divis & Benjamin, 2014). These demonstrations are largely silent on exactly why semantic generation induces context change. One possible explanation is that the retrieval activities involved in semantic generation mediate the induced context change, with the general idea being that a higher degree of retrieval activities induces larger context change. If so, semantic generation may in fact induce more context change than simple calculation or arithmetic tasks do, and substantial context change may also be induced if subjects, for instance, generated autobiographical detail as part of a distractor task. Results from numerous studies are consistent with this view (e.g., Jonker et al., 2013; Klein, Shiffrin, & Criss, 2007; Pastötter & Bäuml, 2007; Sahakyan & Kelley, 2002).

In line with Chan, Manley et al.'s (2018) previous finding, the FTE after retrieval practice in Experiment 1B was found to be accompanied by increased clustering scores when categorized lists were used, thus supporting the assumption that, for this type of material, retrieval practice induces optimization of encoding and retrieval. No such optimization was found in the semantic-generation condition, which is consistent with the two-factor account. Intrusions during list-3 recall were numerically reduced in the retrieval-practice and semantic-generation conditions, relative to their (restudy and distractor) control conditions, but no significant reductions were observed, which contrasts with some prior work (e.g., Chan, Manley, Davis, & Szpunar, 2018; Divis & Benjamin, 2014; Szpunar et al., 2008). This difference between studies may be due to the fact that intrusions were already fairly rare in the two control conditions, so that not much room was left for major reductions in number of intrusions in response to retrieval activity.

Experiments 2a and 2b

Like Experiments 1a and 1b, Experiments 2a and 2b differed with respect to study material only, and are therefore also reported together. In contrast to Experiments 1a and 1b, Experiments 2a and 2b employed retrieval practice and restudy conditions but no semantic generation and distractor conditions. The goal of Experiments 2a and 2b was to examine whether Chan, Manley et al.'s (2018) finding for categorized lists, that the magnitude of the FTE as induced by retrieval practice is largely maintained when the retention interval between study and test of the critical final list is prolonged to 25 min, generalizes to unrelated lists. To this end, participants in Experiments 2a and 2b engaged in a 3-list FTE task, in which, after study of lists 1 and 2, they were either asked to retrieve the immediately preceding list or to restudy the list, analogous to the retrieval-practice and restudy conditions of Experiments 1a and 1b. Unlike in the preceding two experiments, however, study of list 3 was either followed by a 1-min or a 25-min retention interval, after which subjects were instructed to recall as many list-3 items as possible. Duration of retention intervals followed the Chan, Manley et al. (2018) study.

The two-factor account predicts a difference in persistence of the FTE depending on study material. In particular, the FTE should be reduced or be even absent after prolonged retention interval with unrelated lists, but be maintained across retention intervals with categorized lists. Indeed, for unrelated lists, interpolated retrieval should induce mental context change and thus facilitate retention of the last studied target list after short retention interval. However, because beneficial effects of induced context change typically dissipate relatively quickly after study, subjects should barely benefit from retrieval practice after prolonged retention interval. In contrast, for categorized lists, interpolated tests should induce optimization of encoding and retrieval strategies and, as a result, the FTE should still be observed when the retention interval is prolonged. Similarly, higher clustering scores after retrieval practice than restudy should show up with categorized lists after both retention interval conditions.

Method

Participants. In both Experiment 2a and Experiment 2b, 144 students at Regensburg University participated (Experiment 2a: mean age = 22.7 years; Experiment 2b: mean age = 22.8 years), with 36 subjects in each of an experiment's four experimental conditions. Participants took part in the experiment in return for either partial course credit or a compensatory amount of money. All subjects spoke German as their native language.

Material. Experiment 2a used the same unrelated study material as Experiment 1a, and Experiment 2b used the same categorized study

[2] This alternative explanation was suggested to us during the review process.

material as Experiment 1b.

Design and procedure. The experiment had a 2 × 2 design with the between-participants factors of TYPE OF PRACTICE (restudy vs. retrieval practice) and RETENTION INTERVAL (1-min retention interval vs. 25-min retention interval). Like in the restudy and retrieval-practice conditions of Experiments 1a and 1b, participants were asked after study of lists 1 and 2 to either study the just presented list once again (restudy condition) or to recall as many items from the just presented list as possible (retrieval-practice condition). The 1-min and 25-min retention interval conditions differed in that, in the one condition, participants were asked to recall list-3 items after 1 min of backward counting, whereas in the other condition, participants engaged in two additional tasks – 6 min of solving simple arithmetic tasks followed by 18 min of solving Raven's standard progressive matrices (Raven, Raven, & Court, 2000) – before being tested on list 3.

Results of Experiment 2a

List-3 recall

Correct recall. Fig. 2a shows the percentage of correctly recalled list-3 items in the restudy and retrieval-practice conditions, for both the 1-min and the 25-min retention intervals. A 2 x 2 ANOVA with the between-subjects factors of TYPE OF PRACTICE (restudy vs. retrieval practice) and RETENTION INTERVAL (1-min retention interval vs. 25-min retention interval) revealed main effects of TYPE OF PRACTICE, $F(1,140) = 9.809, MSE = .043, p = .002$, partial $\eta^2 = .065$, and RETENTION INTERVAL, $F(1,140) = 25.858, MSE = .043, p < .001$, partial $\eta^2 = .156$, reflecting better recall in the retrieval-practice than the restudy condition (51.5% vs. 40.7%), and better recall after the 1-min than the 25-min retention interval (54.8% vs. 37.3%). Critically, there was also a significant interaction between the two factors, $F(1, 140) = 7.625, MSE = .043, p = .007$, partial $\eta^2 = .052$, suggesting that the magnitude of the FTE varied with retention interval. Indeed, while planned comparisons between the retrieval-practice and restudy conditions showed that there was a reliable FTE when retention interval was short (64.9% vs. 44.7%), $t(70) = 4.757, p < .001, d = 1.121$, there was no such difference when retention interval was long (38.0% vs. 36.7%), $t(70) < 1$, $B_{01} = 5.445$. For the restudy condition, planned comparisons showed no difference in recall between the 1-min and 25-min retention interval (44.7% vs. 36.7%), $t(70) = 1.530, p = .130$, $B_{01} = 2.587$, whereas there was a significant decrease in recall from the 1-min to the 25-min lag in the retrieval-practice condition (64.9% vs. 38.0%), $t(70) = 6.024, p < .001, d = 1.423$.

Intrusions. Table 1 shows the number of intrusions during list-3 recall in the restudy and retrieval-practice conditions, for both the 1-min and the 25-min retention intervals. A 2 x 2 ANOVA with the factors of TYPE OF PRACTICE and RETENTION INTERVAL revealed main effects of TYPE OF PRACTICE, $F(1,140) = 4.311, MSE = 4.525, p = .040$, partial $\eta^2 = .030$, and DELAY, $F(1,140) = 7.306, MSE = 4.525, p = .008$, partial $\eta^2 = .050$, on number of intrusions, suggesting that participants produced fewer intrusions in the retrieval-practice than the restudy condition (0.83 vs. 1.57), and fewer intrusions after the 1-min than the 25-min retention interval (0.72 vs. 1.68). There was no significant interaction between the two factors, $F(1,140) < 1$, $B_{01} = 8.480$.

Recall across lists

Examining recall performance across the three lists in the retrieval-practice condition, a mixed 3 x 2 ANOVA with the within-subjects factor of LIST (lists 1–3) and the between-subjects factor of RETENTION INTERVAL (1-min retention interval vs. 25-min retention interval) revealed a significant interaction between factors, $F(2,140) = 34.774, MSE = 0.011$, $p < .001$, partial $\eta^2 = .332$. Recall performance across lists was dependent upon whether subjects took part in the 1-min or the 25-min retention interval condition. While participants' recall levels were similar across lists after short retention interval (list 1 = 60.3%, list 2 = 60.8%, list 3 = 64.9%), $F(2,70) = 2.535, MSE = 0.009, p = .087$, partial $\eta^2 = .068$, $B_{01} = 5.724$, after long retention interval, participants recalled substantially fewer words from list 3 than from lists 1 and 2 (list 1 = 58.5%, list 2 = 58.6%, list 3 = 38.0%), $F(2,70) = 41.353, MSE = 0.012, p < .001$, partial $\eta^2 = .542$.

Results of Experiment 2b

List-3 recall

Correct recall. Fig. 2b shows the percentage of correctly recalled list-3 items in the restudy and retrieval-practice conditions, for both the 1-min and the 25-min retention intervals. A 2 x 2 ANOVA with the between-subjects factors of TYPE OF PRACTICE (restudy vs. retrieval practice) and RETENTION INTERVAL (1-min retention interval vs. 25-min retention interval) revealed main effects of TYPE OF PRACTICE, $F(1,140) = 13.416, MSE = .040, p < .001$, partial $\eta^2 = .087$, and RETENTION INTERVAL, $F(1, 140) = 17.202, MSE = .040, p < .001$, partial $\eta^2 = .109$, reflecting better recall in the retrieval-practice than the restudy condition (49.7% vs. 37.4%), and better recall after the 1-min than the 25-min retention interval (50.5% vs. 36.6%). Critically, there was no significant interaction between the two factors, $F(1,140) < 1$, $B_{01} = 11.551$, indicating that the FTE did not vary with retention interval.

Intrusions. Table 1 shows the number of intrusions during list-3 recall in the restudy and retrieval-practice conditions, for both the 1-min and 25-min retention intervals. A 2 x 2 ANOVA with the factors of TYPE OF PRACTICE and RETENTION INTERVAL revealed main effects of TYPE OF PRACTICE, $F(1, 140) = 12.733$, $MSE = 6.013$, $p < .001$, partial $\eta^2 = .083$, and RETENTION INTERVAL, $F(1,140) = 14.230, MSE = 6.013, p < .001$, partial $\eta^2 = .092$, suggesting that participants produced fewer intrusions in the retrieval-practice than the restudy condition (1.40 vs. 2.86), and fewer intrusions after the 1-min retention interval than the 25-min retention interval (1.36 vs. 2.90). There was no significant interaction between the factors, $F(1,140) = 3.752, MSE = 6.013, p = .055$, partial $\eta^2 = .026$, $B_{01} = 1.787$.

Clustering in recall. Regarding ARC scores, a 2 x 2 ANOVA with the factors of TYPE OF PRACTICE and RETENTION INTERVAL revealed a main effect of TYPE OF PRACTICE, $F(1,140) = 13.689, MSE = .165, p < .001$, partial $\eta^2 = .089$, reflecting higher ARC scores in the retrieval-practice than restudy condition (.59 vs..34), but no main effect of RETENTION INTERVAL, $F(1, 140) < 1$, $B_{01} = 11.741$, and no interaction between factors, $F(1, 140) < 1$, $B_{01} = 11.925$ (see Table 1).

Recall across lists

We again examined recall performance and ARC scores across the three lists in the retrieval-practice condition. Regarding recall performance, a mixed 3 x 2 ANOVA with the within-subjects factor of LIST (lists 1–3) and the between-subjects factor of RETENTION INTERVAL (1 min vs. 25 min) revealed a significant interaction between factors, $F(2, 140) = 9.535$, $MSE = 0.017, p < .001$, partial $\eta^2 = .120$. While participants' recall performance was similar across lists in the 1-min retention interval condition (list 1 = 61.5%, list 2 = 59.0%, list 3 = 57.1%), $F(2, 70) = 1.028, MSE = 0.017, p = .318$, partial $\eta^2 = .029$, $B_{01} = 25.695$, those in the 25-min retention interval condition recalled fewer words from list 3 than from lists 1 and 2 (list 1 = 63.6%, list 2 = 59.9%, list 3 = 42.3%), $F(2,70) = 27.478, MSE = 0.017, p < .001$, partial $\eta^2 = .440$.

For the ARC analysis, data across the two delay conditions were collapsed, because (i) the two procedures were identical for lists 1 and 2 and (ii) the preceding analyses showed that delay did not affect the clustering of list-3 items in the retrieval-practice conditions (see also Chan et al., 2018). ANOVA showed that ARC scores rose across lists, $F(1, 35) = 5.363$, $MSE = .061, p = .023$, partial $\eta^2 = .070$, with the ARC score increasing from.46 in list 1 to.57 in list 2 and.59 in list 3.

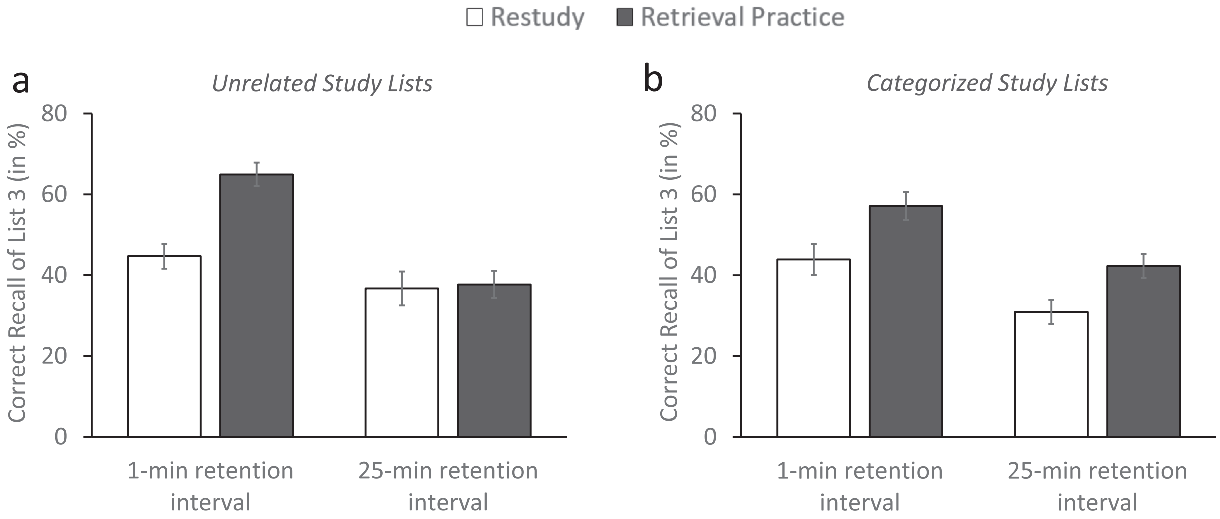

Fig. 2. (a) Results of Experiments 2a. Mean correct list-3 recall as a function of the retention interval between study and test (1-min retention interval, 25-min retention interval) and the type of practice (restudy, retrieval practice) for lists of unrelated items. (b) Results of Experiments 2b. Mean correct list-3 recall as a function of the retention interval between study and test (1-min retention interval, 25-min retention interval) and the type of practice (restudy, retrieval practice) for lists of categorized items. Error bars represent standard errors.

Discussion

Similar to Experiments 1a and 1b, the results of Experiments 2a and 2b show retrieval-practice effects after the 1-min retention interval, with higher recall for the critical list when there was retrieval practice on the previously studied lists than when these lists were restudied. More important, the experiments demonstrate an effect of study material on the persistence of the effect: whereas the FTE maintained in size with longer retention interval with categorized lists, it disappeared with unrelated lists. These results are consistent with the proposed two-factor account of the FTE. This account suggests that mainly strategy change mediates the FTE with categorized lists and thus should create FTEs after both short and long retention interval. In contrast, it suggests that mainly context change mediates the FTE with unrelated lists and thus should show a declining FTE with longer retention interval for this type of list. The present results align with these suggestions. The results of the two experiments again challenge explanations of the FTE that rely solely on one of the two cognitive mechanisms, which would predict that the FTE maintains (strategy-change explanation) or disappears (context-change explanation) with longer retention interval regardless of material.

We replicated the finding that, with categorized lists, the FTE is accompanied by increased clustering scores even when there was a 25-min retention interval (Chan, Manley et al., 2018), thus supporting the assumption that, with this type of material, retrieval practice may induce a strategy change that still facilitates retrieval after prolonged retention interval. The number of intrusions that was produced at the test of the critical final list was lower in the retrieval-practice than the restudy condition even after the prolonged delay, with both unrelated and categorized lists. The results for categorized lists replicate findings of Chan, Manley et al. (2018). For unrelated lists, the observation that intrusions in the retrieval-practice condition were still reduced after prolonged retention interval is surprising because no FTE arose with unrelated lists after the prolonged delay. However, the effect was driven by two atypical subjects,[3] so that future work may not easily replicate this surprising effect.

Experiments 3a and 3b

Experiments 3a and 3b again only differed with respect to study material, and therefore, are reported together. Like Experiments 2a and 2b, the two experiments included retrieval-practice and restudy conditions only. The goal of Experiments 3a and 3b was to examine whether the observation that, with categorized lists, the size of the FTE largely maintains when the learning of the critical list is lagged (Chan, Manley et al., 2018), generalizes to unrelated lists. Like in the previous experiments of this study, a 3-list FTE task was employed in Experiments 3a and 3b. After study of lists 1 and 2, participants were either asked to retrieve the immediately preceding list or to restudy the list. Critically, prior to study of list 3, there was either a 1-min or a 25-min lag. After study of list 3, participants in both lag conditions were asked to recall as many list-3 items as possible after a 1-min retention interval. Lag conditions followed the Chan, Manley et al. study.

The two-factor account predicts different FTEs after prolonged lag depending on study material. In particular, after prolonged lag, the FTE should be reduced or be even absent with unrelated lists. Prolonged lag should induce a context change also in the absence of retrieval practice and thus facilitate the learning of the subsequent critical list. As a result, retrieval practice may not afford any additional recall benefit. In contrast, with categorized lists, interpolated tests should induce optimization of encoding and retrieval strategies and this benefit should still arise when learning of the critical list was lagged; in both lag conditions, the increase in recall levels should be accompanied by an increase in clustering scores.

Method

Participants. In both Experiment 3a and Experiment 3b, 144 students at Regensburg University participated (Experiment 3a: mean age = 22.5 years; Experiment 3b: mean age = 22.5 years), with 36 subjects in each of an experiment's four experimental conditions. Participants took part in the experiment in return for either partial course credit or a compensatory amount of money. All subjects spoke German as their native language.

Material. Experiment 3a used the same unrelated study material as Experiment 1a, and Experiment 3b used the same categorized study material as Experiment 1b.

[3] These two subjects took part in the restudy condition and produced unusally high numbers of intrusions (17 and 15 intrusions, respectively). When removing these subjects from the data set, the difference between the retrieval-practice and restudy conditions disappeared (1.17 vs. 1.38), $t(68) < 1$, $B_{01} = 4.742$.

Design. The experiment had a 2 × 2 design with the between-participants factors of type of practice (restudy vs. retrieval practice) and lag (1-min lag vs. 25-min lag). In the restudy and retrieval-practice conditions, participants were asked after study of lists 1 and 2 to either study the just presented list once again (restudy condition) or to recall as many items from the just presented list as possible (retrieval-practice condition). The 1-min and the 25-min lag conditions differed in that, in the former condition, participants were asked to study list 3 after 1 min of backward counting, whereas in the latter condition, participants engaged in two additional tasks – 6 min of solving simple arithmetic tasks followed by 18 min of solving Raven's standard progressive matrices (Raven et al., 2000) – before studying list 3.

Procedure. The procedure was identical to the 1-min retention interval condition of Experiments 2a and 2b, with the sole exception that either a time interval of 1 min or a time interval of 25 min preceded the study of list 3.

Results of Experiments 3a

List-3 recall

Correct recall. Fig. 3a shows the percentage of correctly recalled list-3 items in the restudy and retrieval-practice conditions, for both the 1-min and the 25-min lags. A 2 x 2 ANOVA with the between-subjects factors of type of practice (restudy vs. retrieval practice) and lag (1-min lag vs. 25-min lag) revealed a main effect of type of practice, $F(1,140) = 15.678$, $MSE = .045, p < .001$, partial $\eta^2 = .101$, reflecting better recall in the retrieval-practice than the restudy condition (63.2% vs. 49.2%), but no main effect of lag, $F(1,140) < 1$, $B_{01} = 10.953$. Critically, there was a significant interaction between the two factors, $F(1, 140) = 7.020$, $MSE = .045, p = .009$, partial $\eta^2 = .048$, suggesting that the magnitude of the FTE varied with lag. Indeed, planned comparisons between the retrieval-practice and restudy conditions showed that while there was a reliable FTE in the short-lag condition (68.6% vs. 45.3%), $t(70) = 4.541, p < .001, d = 1.070$, there was no FTE in the long-lag condition (57.8% vs. 53.1%), $t(70) < 1$, $B_{01} = 5.304$. For the restudy condition, planned comparisons showed no difference in recall between the 1-min and 25-min lag (45.3% vs. 53.1%), $t(70) = 1.321, p = .191$, $B_{01} = 3.511$, whereas there was a significant decrease in recall from the 1-min to the 25-min lag in the retrieval-practice condition (68.6% vs. 57.8%), $t(70) = 2.851, p = .006, d = 0.671$.

Intrusions. Table 1 shows the number of intrusions during list-3 recall in the restudy and retrieval-practice conditions, for both the 1-min and 25-min lags. A 2 x 2 ANOVA with the factors of type of practice and lag revealed no main effects of type of practice, $F(1,140) < 1$, $B_{01} = 11.944$, or lag, $F(1,140) = 2.047, MSE = 0.763, p = .155$, partial $\eta^2 = .014$, $B_{01} = 4.218$, and also no interaction between factors, $F(1, 140) = 2.047$, $MSE = 0.763, p = .155$, partial $\eta^2 = .014$, $B_{01} = 4.218$.

Recall across lists

Examining recall performance across the three lists in the retrieval-practice condition, a mixed 3 x 2 ANOVA with the within-subjects factor of list (lists 1–3) and the between-subjects factor of lag (1-min lag vs. 25-min lag) revealed no main effects of list, $F(2, 140) < 1$, $B_{01} =$ 52,631, or lag, $F(1,70) = 3.791, MSE = 0.060, p = .056$, partial $\eta^2 = .051$, $B_{01} = 1.268$, and no interaction between factors, $F(1,70) = 2.468$, $MSE = 0.011, p = .088$, partial $\eta^2 = .034$, $B_{01} = 20.890$. Averaged across the two lags, the percentage of correctly recalled items from lists 1 to 3 was 62.1%, 64.0%, and 63.2%.

Results of Experiment 3b

List-3 recall

Correct recall. Fig. 3b shows the percentage of correctly recalled list-3 items in the restudy and retrieval-practice conditions, for both the 1-min and the 25-min lags. A 2 x 2 ANOVA with the between-subjects factors of type of practice (restudy vs. retrieval practice) and lag (1-min lag vs. 25-min lag) revealed a main effect of type of practice, $F(1,140) = 19.203$, $MSE = .030, p < .001$, partial $\eta^2 = .121$, reflecting better recall in the retrieval-practice than the restudy condition (60.9% vs. 48.3%), but no main effect of lag, $F(1,140) < 1$, $B_{01} = 11.978$, and no interaction between the two factors, $F(1,140) < 1$, $B_{01} = 11.938$. The FTE therefore did not vary with lag.

Intrusions. Table 1 shows the number of intrusions during list-3 recall in the restudy and retrieval-practice conditions, for both the 1-min and 25-min lags. A 2 x 2 ANOVA with the factors of type of practice and lag revealed no main effects of type of practice, $F(1,140) < 1$, $B_{01} = 11.950$, or lag, $F(1, 140) < 1$, $B_{01} = 8.612$, and also no significant interaction between factors, $F(1,140) = 2.205, MSE = 0.871, p = .131$, partial $\eta^2 = .016$, $B_{01} = 3.704$.

Clustering in recall. Regarding ARC scores, a 2 x 2 ANOVA with the factors of type of practice and lag revealed a main effect of type of practice, $F(1, 140) = 5.836, MSE = .159, p = .017$, partial $\eta^2 = .040$, reflecting higher ARC scores in the retrieval-practice than restudy condition (.60 vs..44), but no main effect of lag, $F(1, 140) < 1$, $B_{01} = 9.363$, and no interaction between factors, $F(1,140) < 1$, $B_{01} = 10.883$ (see Table 1).

Recall across lists

We again examined recall performance across the three lists in the retrieval-practice condition. A mixed 3 x 2 ANOVA with the within-subjects factor of list (lists 1–3) and the between-subjects factor of lag (1 min vs. 25 min) revealed a main effect of list, $F(2,70) = 4.495, MSE = 0.011$, $p = .013$, partial $\eta^2 = .060$, but no main effect of lag, $F(1, 70) < 1$, $B_{01} = 8.173$, and also no interaction between factors, $F(2, 70) < 1$, $B_{01} = 48.800$. While recall rates across lists varied statistically, numerically these differences were rather small and unsystematic. Averaged across the two lag conditions, the percentage of correctly recalled items from lists 1 to 3 was 63.2%, 57.9%, and 60.9%.

We also examined ARC clustering scores for participants in the retrieval-practice condition. For this analysis, data were again collapsed across lags, because (i) the two procedures were identical for lists 1 and 2 and (ii) the preceding analyses showed that lag did not affect the clustering of list-3 items in the retrieval-practice condition (see also Chan, Manley et al., 2018). ANOVA showed that ARC scores rose across lists, $F(1,35) = 7.156, MSE = .052, p = .009$, partial $\eta^2 = .092$, with the ARC score increasing from.51 in list 1 to.66 in list 2 and.60 in list 3.

Discussion

Similar to the four previous experiments, the results of Experiments 3a and 3b show retrieval-practice effects after the 1-min lag, with higher recall for the critical list when there was retrieval practice on the previously studied lists than when these lists were restudied. More important, the experiments demonstrate an effect of study material on the persistence of the effect: whereas the effect maintained in size with categorized lists when learning of the critical list was lagged, the effect disappeared with unrelated lists. These results align well with the proposed two-factor account and its indication that, with categorized lists, mainly strategy change mediates the FTE and, with unrelated lists, mainly context change mediates the effect. Accordingly, retrieval practice should create a FTE after both short and long lag with categorized lists, but should induce a FTE after short lag only with unrelated lists, which is exactly what the present results show. The results again pose a problem for accounts of the FTE, which assume that the FTE is mediated by a single mechanism, and which assume that lag influences the FTE regardless of material.

We replicated the finding that, with categorized lists, the FTE is accompanied by increased clustering scores even when there is a 25-min lag (Chan, Manley et al., 2018). This supports the assumption of the two-factor account that, with this type of material, retrieval practice induces a strategy change that still facilitates retrieval after prolonged lag be-

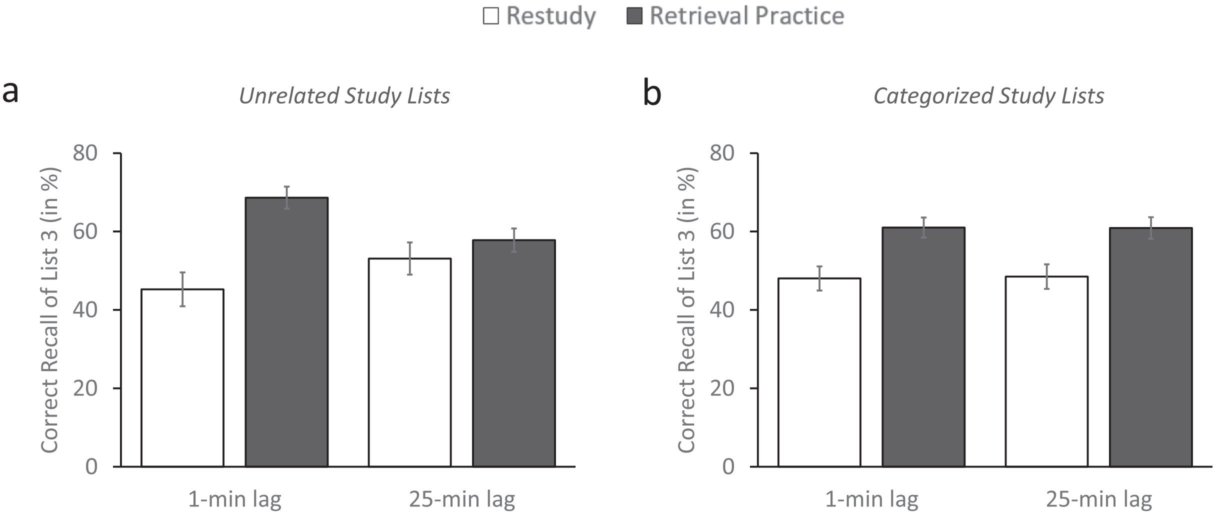

Fig. 3. (a) Results of Experiments 3a. Mean correct list-3 recall as a function of the lag between study of list 3 and the study of the prior lists (1-min lag, 25-min lag) and the type of practice (restudy, retrieval practice) for lists of unrelated items. (b) Results of Experiments 3b. Mean correct list-3 recall as a function of the lag between study of list 3 and the study of the prior lists (1-min lag, 25-min lag) for lists of categorized items. Error bars represent standard errors.

tween the initially studied lists and the critical list. Both in Experiment 3a and Experiment 3b, there was no statistical difference between the retrieval-practice and restudy conditions in number of intrusions. Similar to Experiments 1a and 1b, this is likely due to the fact that intrusions were already fairly rare in the restudy conditions, so that no room was left for a major reduction in intrusions in the retrieval-practice conditions.[4]

One finding of Experiment 3a that warrants further discussion is the observation that in the retrieval-practice condition, list-3 recall was significantly higher after the 1-min lag than the 25-min lag (68.6% vs. 57.8%). This may seem surprising because, a priori, one might expect comparable recall rates for the 1-min and 25-min lag conditions. In fact, prolonged lag should maintain, or even strengthen, the context change induced by the preceding retrieval practice and thus promote list-3 recall. Notice, however, that recall levels of lists 1 and 2 during initial retrieval practice were also (a bit) higher in the 1-min-lag condition (list 1 = 64.0%; list 2 = 66.4%; list 3 = 68.6%) than the 25-min-lag condition (list 1 = 60.2%; list 2 = 61.6%; list 3 = 57.8%), with relatively stable recall levels across lists 1 to 3 within each of the two lag conditions. This pattern suggests that the subject samples differed slightly in performance for the two lag conditions, with better performing subjects in the 1-min lag than the 25-min lag condition – even though differences between lag conditions were not statistically significant (see Results section).

General Discussion

Across six experiments, the present study once again demonstrated that retrieval practice of previously learned material can enhance memory of newly learned material more effectively than does restudy of the previously learned material. This FTE arose for both unrelated lists (Experiments 1a, 2a, 3a) and categorized lists (Experiments 1b, 2b, 3b), which replicates prior work (e.g., Chan, Manley et al., 2018; Pastötter et al., 2011; Szpunar et al., 2008). Critically, however, parallel FTEs for unrelated and categorized study material arose only for short retention interval between study and test of the critical list and short lag prior to study of the critical list. When there was a prolonged retention interval or a prolonged lag, the FTE was observed with categorized lists but disappeared with unrelated lists. As another difference between materials, semantic generation produced an FTE with unrelated lists, but not with categorized lists. Together, these findings demonstrate a critical role of study material for the FTE.

Implications for theoretical accounts of the FTE

The present results are consistent with the suggested two-factor account of the FTE. According to this account, in general, both context change and strategy change can contribute to the FTE. First, the account assumes that retrieval activities, like retrieval practice or semantic generation, induce context change and thus can segregate the single lists and improve recall of the final list. Critically, such context-change effects should take effect mainly with unrelated lists and much less with categorized lists. The reason is that, with categorized lists, study of a list may lead to reinstatement of the previous list's study context - due to the repetition of the same categories as they were employed in the previous lists (Jonker et al., 2013; Wirth & Bäuml, 2020) - and thus eliminate the context-change effect. In contrast, with unrelated lists, such context reinstatement should be rare - because items from different lists typically do not share category membership - and the induced context change should therefore be effective. Second, the account further assumes that retrieval practice can also induce strategy change, because retrieval practice can provide information on further tests and thus lead to optimization of encoding and retrieval strategies. Such strategy change should primarily come into play when categorized material is studied and play a minor role for unrelated lists, which often do not lend themselves to relational processing. In particular, strategy change should be retrieval-practice specific, and not arise in response to semantic generation. Retrieval practice, but not semantic generation, can provide information about the learning task at hand.

According to the two-factor account, the FTE should therefore be

[4] In all conditions of the present experiments in which retention interval and lag were short (1 min), number of intrusions was numerically lower in the retrieval-practice than the restudy condition. This holds while statistical power was often insufficient to demonstrate statistical significance of the difference. To increase statistical power, we collapsed the data of Experiments 1A, 2A, and 3A – in each of which unrelated lists were applied — and of Experiments 1B, 2B, and 3B – in each of which categorized lists were applied. All subjects were included who engaged in (i) the retrieval-practice and restudy conditions of Experiments 1a and 1b, (ii) the 1-min retention interval condition of Experiments 2a and 2b, and (iii) the 1-min lag condition of Experiments 3a and 3b. Paired comparisons showed that intrusion rates were significantly lower in the retrieval-practice than the restudy conditions, for both unrelated lists (0.44 vs. 0.74), $t(214) = 2.327, p = .021, d = 0.317$, and categorized lists (0.94 vs. 1.49), $t(70) = 3.018, p = .003, d = 0.411$.

primarily triggered by context change with unrelated lists and by strategy change with categorized lists. As a result, the FTE with unrelated lists should not be retrieval-practice specific and be dependent on short retention interval and short lag prior to study of the critical list. The dependence on short retention interval should arise because context-change effects have been shown to be transient in nature (Abel & Bäuml, 2017; Divis & Benjamin, 2014), and the dependence on short lag should arise because a prolonged lag typically induces context change also in the absence of retrieval practice (Estes, 1955; Mensink & Raaijmakers, 1988). In contrast, the FTE with categorized lists should be retrieval-practice specific, be less dependent on retention interval, and still be present after prolonged lag. The reduced dependencies on retention interval and lag should be the result of a strategy change that still transpires after prolonged retention interval and lag. The pattern of results observed across the present six experiments agrees with these predictions.

Whereas the present results support the two-factor explanation, they are inconsistent with single-mechanism views of the FTE as they are reflected in the strategy-change and context-change accounts. The strategy-change explanation predicts quantitatively different FTEs for categorized and unrelated lists, with the sole difference being that the FTE should be reduced in magnitude with unrelated lists relative to categorized lists. Indeed, unrelated lists should leave less room for strategy change than categorized lists because unrelated material is less suited for relational processing (see Chan, Manley et al., 2018). For both types of study material, the FTE should arise after prolonged retention interval and prolonged lag, and for both types of study material, the FTE should show retrieval-practice specificity and not show up in response to semantic generation. The present results do not align with these predictions and also provide no indication that the magnitude of the FTE is more pronounced with categorized lists.[5] The present results are also inconsistent with the context-change explanation of the FTE, which suggests FTEs after both retrieval practice and semantic generation and indicates that the FTE should disappear after both prolonged retention interval and prolonged lag. In particular, the account does not propose a role of study material for the FTE, which contrasts with our findings.

Although the suggested two-factor account is able to explain the present results, open questions remain. Regarding the strategy change component, for instance, the question arises of whether the suggested strategy change takes effect primarily at the encoding or the retrieval stage, or whether some form of interaction between encoding and retrieval change induces the FTE. Future research should also establish whether the observed increase in semantic clustering following retrieval practice is the result of metacognitive monitoring and high-level control processes, or is due to processes that do not involve conscious adjustments in encoding or retrieval strategy. Regarding the context component, an open question is why retrieval practice and semantic generation should enhance context change relative to simple calculation or mental arithmetic tasks. Here we suggest that retrieval activities induce context change, assuming that retrieval practice and semantic generation include more retrieval activities and thus induce more context change than simple calculation or arithmetic tasks do (see also Discussion of Experiment 1 above). While this assumption is able to explain the present results – as it has been able to explain results in other paradigms, like the list-before-last paradigm (Jang & Huber, 2008) or the retrieval-induced forgetting paradigm (Jonker et al., 2013) – it is silent on exactly how retrieval causes context change and why it may induce more context change than tasks involving less retrieval activities. It is a high priority for future work to address this thorny issue more directly. As a corollary, such work might also provide a method to examine a priori whether a given task would, or would not, induce internal context change, which would exclude possible circularity problems in any context change explanations (see Chan, Manley, et al., 2018).

Relation to prior FTE studies

The present findings are consistent with other findings from the FTE literature and may also help to reconcile apparently conflicting results. For instance, the present results align with prior work in which the FTE has been found not only in response to retrieval practice but also in response to semantic generation. This result has been reported for unrelated lists (Divis & Benjamin, 2014; Pastötter et al., 2011). The present study replicates the finding and demonstrates that it does not generalize to categorized lists. The present study also replicates the Chan, Manley et al. (2018) finding that the FTE maintains across prolonged retention interval and prolonged lag with categorized lists, but shows that the FTE does not persist across prolonged retention interval or lag with unrelated lists. These results help reconciling two conflicting lines of findings in the prior work: while Chan, Manley et al. found the FTE to persist across prolonged retention interval, Divis and Benjamin (2014) found the effect to be quite transient. The present study solves the conflict by showing that the role of retention interval for the FTE varies with material: the effect lasts with categorized lists, as they were employed in Chan, Manley et al.'s (2018) study, but is more transient with unrelated lists, as they were employed in Divis and Benjamin's (2014) study. All of these findings are in line with the two-factor account.

While the present two-factor account suggests critical roles of strategy change and context change for the FTE, prior work has proposed other multi-factor accounts (see Chan, Meissner et al., 2018). For instance, Chan, Manley, and Ahn (2020) recently argued that retrieval practice of initially studied (nontarget) material can enhance participants' encoding and retrieval strategies *and* improve their attention to the newly studied material. In line with this suggestion, these researchers found that for categorized study lists, test expectancy ratings – which were used as a proxy for attention – and semantic clustering scores – which were used as a proxy for strategy use – were jointly but independently predictive of the size of the FTE. However, similar to the strategy-change explanation, this two-factor explanation cannot easily account for the present finding that, with unrelated lists, an FTE arose after semantic generation (Experiment 1B), but no FTEs were observed after prolonged lag (Experiment 2B) and prolonged retention interval (Experiment 3B). Indeed, if semantic generation led to an FTE for unrelated lists because the generation task released attentional resources, then one should expect a comparable FTE for categorized lists in Experiment 1B, which was not the case. Similarly, if strategy change and attention restoration mediated the FTE in the present experiments, maintained FTE should have been observed after prolonged lag and prolonged retention interval for both categorized and unrelated lists, and not merely for categorized lists, as the results of Experiments 2 and 3 suggest.

Other researchers have also pointed to a potential role of attentional factors for the FTE. In fact, in several recent studies, the proposal has been made that attentional resources may decline with the study of several lists and that retrieval practice may induce a reset of the attentional processes and thus improve the encoding of the new items (Jing, Szpunar, & Schacter, 2016; Pastötter et al., 2011). Evidence for this view comes from both electrophysiological and behavioral research. For instance, Pastötter et al. (2011) measured brain oscillations while subjects studied five unrelated lists of words either with or without retrieval

[5] In fact, we collapsed the data of the three experiments using unrelated lists (Experiments 1a, 2a, 3a) and the three experiments using categorized lists (Experiments 1b, 2b, 3b), and included those subjects who engaged in (i) the retrieval-practice and restudy conditions of Experiments 1a and 1b, (ii) the 1-min retention interval condition of Experiments 2a and 2b, and (iii) the 1-min lag condition of Experiments 3a and 3b. Doing so, we found that, numerically, the FTE was even slightly more pronounced for unrelated than categorized lists (19.4% vs. 14.4%). Notice, however, that while the magnitude of the FTE may vary between unrelated and categorized lists due to the varying degree of relatedness of the study lists, other factors, like the mean frequency or concreteness of the single list items, may influence the size of the effect as well.

practice after study of each of the first four lists. The results not only showed a typical FTE for the critical list 5, but also indicated that alpha band activity (8–14 Hz), which has been linked to memory load (Jensen, Gelfand, Kounios, & Lisman, 2002) and inattention (Palva & Palva, 2007), rose during study of the five lists in the absence of retrieval practice. In contrast, no increase in alpha power was observed in the presence of retrieval practice. Additional, behavioral evidence for the view, for instance, comes from Szpunar et al. (2013), who had their subjects watch four video segments about an introductory course to statistics, either with or without retrieval practice after presentation of the first three segments. Results not only showed a typical FTE for the critical fourth video segment but also indicated that interpolated retrieval practice enabled participants to maintain a high level of attention to encoding by increasing their note-taking and reducing their mind wandering. Taken together, these studies suggest that retrieval practice may induce a reset of attentional processes, both when subjects study unrelated material (Pastötter et al., 2011) and when they study more related materials (Szpunar et al., 2013).

The FTE with complex study material

The present and prior results suggest that the two-factor account may explain the FTE with word lists, but it is less clear whether it may also explain the FTE with more complex study material, like, for instance, prose passages. Thus far, only few studies have examined the FTE in the learning of complex texts (for a review, see Yang, Potts, & Shanks, 2018). One of these prior studies, for instance, used a prose passage about the U.S. labor market, which was segmented into three sections, and found an FTE for the critical third section when subjects were asked to retrieve Sections 1 and 2 following initial study (Wissman et al., 2011). Using a similar approach, Jing et al. (2016) segmented a lecture video on public health into eight 5-min sections, and found an FTE for the critical last section of the video (for related results, see Szpunar et al., 2013). Divis and Benjamin (2014) showed that the FTE in the learning of prose passages can also show up when there is interpolated semantic generation between the single sections, thus suggesting that the FTE may not be specific to retrieval practice with complex material.

While the Divis and Benjamin (2014) finding may suggest a role of context change for the FTE with more complex material, drawing firm conclusions on the basis of the current results may be premature. Indeed, further work is required that, for instance, examines persistence of the effect with this type of material as well as dependence of the effect on lag between initially studied prose material and subsequently studied (target) prose material. Such work would help to indicate whether context change or strategy change may contribute to the effect. Moreover, prose materials can differ in a number of ways – for instance, be more or less coherent thematically – and it is unclear whether the same mechanisms contribute to the FTE with different prose materials. Examining whether the same, or at least similar, cognitive mechanisms contribute to the FTE with word lists and prose material is a high priority for future work on the effect.

Conclusions

In a series of six experiments, we showed a critical influence of study material for the FTE. With categorized lists, the FTE was retrieval-practice specific and still present after prolonged retention interval and lag. In contrast, with unrelated lists, the FTE generalized to semantic generation and disappeared with both prolonged retention interval and lag. These findings are consistent with a new two-factor explanation of the FTE, which assumes contributions of both strategy change and context change for the FTE. This account suggests that, with categorized material, the FTE is mainly driven by strategy change and, with unrelated material, it is mainly driven by context change.

Author Notes

This work was supported by a grant from the German Research Foundation (Deutsche Forschungsgemeinschaft, DFG) awarded to Karl-Heinz T. Bäuml (BA 1382/18–1).

References

Abel, M., & Bäuml, K.-H. T. (2017). Testing the context-change account of list-method directed forgetting: The role of retention interval. *Journal of Memory and Language, 92*, 170–182. https://doi.org/10.1016/j.jml.2016.06.009.

Abel, M., & Bäuml, K.-H. T. (2019). List-method directed forgetting after prolonged retention interval: Further challenges to contemporary accounts. *Journal of Memory and Language, 106*, 18–28. https://doi.org/10.1016/j.jml.2019.02.002.

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. https://doi.org/10.1016/j.jml.2012.07.006.

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. https://doi.org/10.1037/bul0000166.

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. https://doi.org/10.1080/17470218.2016.1175485.

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.

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.

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. https://doi.org/10.1037/0278-7393.34.1.112.

Jensen, O., Gelfand, J., Kounios, J., & Lisman, J. E. (2002). Oscillations in the alpha band (9–12 Hz) increase with memory load during retention in a short-term memory task. *Cerebral Cortex, 12*(8), 877–882. https://doi.org/10.1093/cercor/12.8.877.

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. https://doi.org/10.1037/xap0000087.

Jonker, T. R., Seli, P., & MacLeod, C. M. (2013). Putting retrieval-induced forgetting into context: an inhibition-free, context-based account. *Psychological Review, 120*(4), 852–872. https://doi.org/10.1037/a0034246.

Klein, K. A., Shiffrin, R. M., & Criss, A. H. (2007). Putting context in context. In J. S. Nairne (Ed.), *The foundations of remembering: Essays in honor of Henry L. Roediger III*. New York: Psychology Press.

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.

Palva, S., & Palva, J. M. (2007). New vistas for alpha-frequency band oscillations. *Trends in Neurosciences, 30*(4), 150–158. https://doi.org/10.1016/j.tins.2007.02.001.

Pastötter, B., & Bäuml, K. H. (2007). The crucial role of postcue encoding in directed forgetting and context-dependent forgetting. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 33*(5), 977–982. https://doi.org/10.1037/0278-7393.33.5.977.

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

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

Raven, J., Raven, J. C., & Court, J. H. (2000). *Standard progressive matrices*. Oxford: Psychology Press.

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

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

Roenker, D. L., Thompson, C. P., & Brown, S. C. (1971). Comparison of measures for the estimation of clustering in free recall. *Psychological Bulletin, 76*(1), 45–48. https://doi.org/10.1037/h0031355.

Rupprecht, J., & Bäuml, K.-H. T. (2017). Retrieval-induced versus context-induced forgetting: Can restudy preceded by context change simulate retrieval-induced forgetting? *Journal of Memory and Language, 93*, 259–275.

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. https://doi.org/10.1037/0278-7393.28.6.1064.

Shiffrin, R. M. (1970). Forgetting: Trace erosion or retrieval failure? *Science, 168*(3939), 1601–1603. https://doi.org/10.1126/science.168.3939.1601.

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.

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. (2015). Not all retrieval during learning facilitates subsequent memory encoding [Conference presentation]. Annual meeting of the Psychonomic Society, Chicago, IL, United States.

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.

Wirth, M., & Bäuml, K.-H. T. (2020). Category labels can influence the effects of selective retrieval on nonretrieved items. *Memory and Cognition, 48*, 481–493. https://doi.org/10.3758/s13421-019-00984-8.

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.

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. https://doi.org/10.1038/s41539-018-0024-y.