Journal of Memory and Language 93 (2017) 259–275

Contents lists available at ScienceDirect

Journal of Memory and Language

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

Retrieval-induced versus context-induced forgetting: Can restudy preceded by context change simulate retrieval-induced forgetting?

Julia Rupprecht, Karl-Heinz T. Bäuml [*]

Department of Experimental Psychology, Regensburg University, Germany

ARTICLE INFO

Article history:
Received 21 June 2016
revision received 26 October 2016

Keywords:
Episodic memory
Retrieval-induced forgetting
Context change
Inhibition

ABSTRACT

Retrieval-induced forgetting (RIF) refers to the finding that retrieval practice on a subset of studied items can induce later forgetting of related unpracticed items. The context account of RIF, which attributes RIF to a mismatch of study context and reinstated context at test for the unpracticed items, claims that RIF effects can be simulated by restudy trials when these trials are preceded by context change. To test this proposal, we compared across three experiments effects of retrieval practice and of restudy trials preceded by context change, employing both recall and item recognition testing. We found retrieval practice to impair both recall and recognition of unpracticed items, which is consistent with prior work. In contrast, restudy preceded by context change impaired recall but not recognition of the items. These findings suggest that restudy preceded by context change cannot simulate RIF, which challenges the context account of RIF. The results are consistent with the view of a critical role of retrieval and inhibition in RIF.

Introduction

Retrieval-induced forgetting (RIF) refers to the finding that active retrieval of a subset of previously studied information impairs memory for related nonretrieved information (Anderson, Bjork, & Bjork, 1994). To investigate RIF, a paradigm with three main phases is typically employed: a study phase, in which participants study category-exemplar pairs (e.g., BIRD - *chicken*, SPICE - *ginger*, SPICE - *salt*, etc.); a practice phase, in which half of the exemplars from half of the categories are repeatedly retrieved (e.g., SPICE - *gi_*); and a final test phase, in which all previously studied items are tested using a cued recall test (e.g., BIRD - *c_*, SPICE - *s_*, SPICE - *g_*). The typical finding is that recall of the practiced items (*ginger*) is enhanced and recall of the unpracticed items from the practiced categories (*salt*) is reduced when compared to recall of the control items from the utterly unpracticed categories (*chicken*). The forgetting of the unpracticed items has been proven to be a very robust finding and to prevail over a wide range of materials, settings, and memory tests (for reviews, see Anderson, 2003; Bäuml, Pastötter, & Hanslmayr, 2010; Storm & Levy, 2012).

The two most prominent accounts of RIF attribute the forgetting of the unpracticed items to either inhibition or blocking. Proponents of the inhibition account assume that, during practice, the not-to-be-practiced category exemplars interfere and are actively inhibited to reduce the interference. The inhibitory effect is supposed to be long-lasting and, thus, to manifest itself in the impaired recall of the unpracticed items on the final memory test (e.g., Anderson et al., 1994; Anderson & Spellman, 1995). In contrast, proponents of the blocking account assume that the cue-item associations of the practiced items are

[*] Corresponding author at: Department of Experimental Psychology, Regensburg University, 93040 Regensburg, Germany.
E-mail address: karl-heinz.baeuml@ur.de (K.-H.T. Bäuml).

http://dx.doi.org/10.1016/j.jml.2016.10.006
0749-596X/© 2016 Elsevier Inc. All rights reserved.

strengthened during practice and such strengthening introduces interference of these items during recall of the unpracticed items, thus leading to blocking and impaired recall of the unpracticed items (Raaijmakers & Jakab, 2012; Verde, 2013). Both inhibition and blocking have been argued to account for a wide range of RIF findings (e.g., Anderson, 2003; Raaijmakers & Jakab, 2013), although each of the two accounts is challenged by at least some RIF findings (see section 'General discussion').

More recently, a new account of RIF has been suggested, attributing the forgetting of the unpracticed items to contextual change (Jonker, Seli, & MacLeod, 2013). According to this view, during the practice phase, the act of retrieval induces a change in context and thus generates two distinct contexts for the study and practice phases. In the final test phase, the category labels of the control items from the unpracticed categories (e.g., BIRD) are assumed to trigger reactivation of the study context, which is the only associated context for these items; in contrast, the category labels of the items from the practiced categories (e.g., SPICE) may reactivate the practice context, because it is the more recent context for these categories and the practiced items have been elaborated herein. Thus, for the practiced items (*ginger*) and the control items (*chicken*) an appropriate context may be reinstated, whereas for the unpracticed items from the practiced categories (*salt*), which, like the control items, were present in the study phase only, an inappropriate context may be accessed. The resulting contextual mismatch for the unpracticed items is supposed to underlie the forgetting of the items and to be at the heart of the RIF effect.

Evidence in favor of the context account of RIF

Several findings from the literature indeed corroborate the general idea that retrieval can promote context change. For instance, Szpunar, McDermott, and Roediger (2008) found evidence in multiple-list learning that interpolation of retrieval practice between the study of single lists can lead to higher recall rates and fewer prior-list intrusions for a finally studied target list. This result has been interpreted in terms of reduced proactive interference, arguing that retrieval may enhance list isolation, possibly by inducing context change (e.g., Bäuml & Kliegl, 2013; Pastötter, Schicker, Niedernhuber, & Bäuml, 2011). Similar results arose when between the study of the single lists semantic generation tasks were interpolated, in which subjects were asked to generate as many exemplars of a given nonstudied category as possible. Again, recall performance of the final target list was improved, but in addition, recall of the initially studied first list was impaired (Divis & Benjamin, 2014; Pastötter et al., 2011). This pattern of results resembles the one found in previous work on context-dependent forgetting, in which subjects were instructed to change their internal context between the study of lists (e.g., Pastötter & Bäuml, 2007; Sahakyan & Kelley, 2002), indicating that (semantic) retrieval may indeed drive contextual change (for related results, see Jang & Huber, 2008; Sahakyan & Hendricks, 2012; Shiffrin, 1970).

More specific evidence in favor of the context account of RIF comes from studies showing that RIF does not only arise from competitive but also from noncompetitive retrieval practice. Competitive retrieval practice refers to the standard retrieval practice condition, in which a studied item itself is selectively retrieved (*ginger*) facing interference from other studied category exemplars (*salt*). Extending previous work by Anderson, Bjork, and Bjork (2000), Raaijmakers and Jakab (2012) examined the effects of noncompetitive retrieval practice and showed that retrieving the category label of a to-be-practiced item without retrieval of the to-be-practiced item itself (e.g., _ - *ginger*) can already be sufficient to induce RIF (see also Grundgeiger, 2014). Similarly, Jonker and MacLeod (2012) asked subjects to study category-item pairs (e.g., PET - *dog*) but replaced standard (competitive) retrieval practice with subordinate generation (e.g., *dog* - _), in which the exemplar was presented intact and subjects were instructed to generate a type of *dog*, such as *beagle*. Again, RIF-like forgetting arose. These findings cannot easily be attributed to inhibition, because noncompetitive practice should not induce interference from other category exemplars and thus should not induce inhibition. The findings, however, are consistent with the context account of RIF, because both competitive and noncompetitive retrieval may produce context change and thus induce RIF.

Particularly relevant for the context account of RIF are recent experiments, in which Jonker et al., 2013 tested a core assumption of the context account directly. The rationale of the experiments was that if RIF represents context-dependent forgetting, then one should be able to simulate RIF using restudy for practice and a preceding context change manipulation. Jonker et al. (2013) let participants study category-exemplar pairs, restudy a subset of the category-exemplar pairs from a subset of the categories, and finally tested the whole study list employing a cued recall test. Whereas in one experiment, no context change was induced before practice, in two other experiments, an imagination task was interspersed immediately before the practice phase to change subjects' internal context. In one of the two imagination experiments, subjects were also asked to mentally reinstate the study context immediately before the final test started. The context account predicts that restudy induces (i) no RIF-like forgetting when no preceding context change occurs, (ii) RIF-like forgetting when a context change has been induced, and (iii) no RIF-like forgetting when the study context is reinstated before test. The results confirmed all three predictions, indicating that retrieval may not be necessary for RIF and rather context change followed by selective restudy may be sufficient to induce forgetting of the unpracticed items (for further results, see Jonker et al., 2013, and section 'General discussion').

From recall to recognition testing

A core assumption of the inhibition account is that RIF is retrieval specific, i.e., it arises following retrieval but not following restudy trials (e.g., Anderson, 2003). And, indeed, comparing the effects of retrieval practice with the effects of standard restudy cycles, the results of numerous studies consistently showed that retrieval practice, but not restudy, induces forgetting of related items (e.g., Bäuml, 2002; Ciranni & Shimamura, 1999; Shivde & Anderson, 2001). In contrast, according to the context account, RIF should not be retrieval specific, but arise due to context change. Importantly, whereas retrieval practice may induce such context change (Jang & Huber, 2008; Shiffrin, 1970), context change may also be prompted by other manipulations, like imagination tasks (Pastötter & Bäuml, 2007; Sahakyan & Kelley, 2002) or semantic generation (Divis & Benjamin, 2014; Jang & Huber, 2008; Pastötter et al., 2011). Thus, the context account predicts that restudy should induce RIF-like forgetting similar to how retrieval practice induces RIF if some context change preceded the restudy trials. Jonker et al.'s (2013) finding, that restudy can produce RIF-like forgetting in recall when an imagination task precedes the practice phase, supports this proposal and suggests that, under such condition, restudy can simulate the effects to retrieval practice trials. However, concluding from this result that RIF generally reflects a context change effect may be premature. Indeed, before drawing firm conclusions on the issue, it needs to be shown that the results reported by Jonker et al. (2013) are not restricted to recall testing but generalize to other testing formats.

Indeed, RIF is not a pure recall phenomenon but can also be found in other testing formats, like, for instance, item recognition (Gómez-Ariza, Lechuga, Pelegrina, & Bajo, 2005; Hicks & Starns, 2004; Spitzer & Bäuml, 2007; Veling & van Knippenberg, 2004; see also Spitzer, 2014, and the results of the large-scale meta-analysis by Murayama, Miyatsu, Buchli, & Storm, 2014; for a demonstration of RIF in forced choice recognition, see Wimber, Alink, Charest, Kriegeskorte, & Anderson, 2015). The finding of reduced recognition of the unpracticed items in response to retrieval practice is consistent with the inhibition account of RIF, because inhibition is supposed to impair unpracticed items' memory representations, which should reduce both recall and recognition of the items. It is less clear a priori whether the finding is also consistent with the context account of RIF. Indeed, whereas some context change studies reported evidence for context effects in item recognition (e.g., Bodner & Lindsay, 2003; Bodner & Richardson-Champion, 2007; Craik & Schloerscheidt, 2011), other studies did not (e.g., Godden & Baddeley, 1975, 1980; Smith, Glenberg, & Bjork, 1978), indicating that, in general, the circumstances that surround a context change may determine whether context change reduces item recognition.

However, for the standard retrieval-practice paradigm, the context account predicts that RIF should generalize from recall to item recognition. The rationale is that, in this paradigm, participants use context information when an exemplar is presented during a recognition test - at least if the judgments are not speeded and the distractors are difficult to distinguish from the targets, which is assumed to be the case when semantically categorized lists are employed as study material (see Jonker et al., 2013, p. 868). Thus, when the studied item *salt* is presented at test without its category label, participants may covertly retrieve the category label SPICE and make use of the context associated with that category in their retrieval attempt. According to the context account, the most relevant and accessible context in the case of practiced categories will then be the practice context, which will prevent context reinstatement for the unpracticed items and thus induce RIF.

While the context account is thus consistent with the finding of RIF in item recognition, critically, it additionally predicts reduced recognition of unpracticed items after restudy cycles when restudy is preceded by context change. Indeed, because, according to the context account, the effects of restudy when preceded by a context change manipulation should be equivalent to the effects of retrieval practice, both retrieval practice and restudy should induce forgetting of unpracticed items, in both recall and item recognition. To the best of our knowledge, no study has yet examined this core prediction of the context account. The present study addresses the issue in three experiments, measuring cued recall and item recognition of the unpracticed items following standard retrieval practice as well as restudy trials preceded by context change.

Overview of experiments

To investigate possible detrimental effects of restudy when preceded by context change on recall and recognition of unpracticed items, we used an imagination task to change participants' internal context in one experiment and a semantic generation task in another. Both tasks have been shown in prior work to accelerate context fluctuation and induce context-dependent forgetting of previously studied items (e.g., Divis & Benjamin, 2014; Pastötter & Bäuml, 2007; Pastötter et al., 2011; Sahakyan & Kelley, 2002). Because, in general, the circumstances that surround a context change may determine whether context change reduces item recognition (see above), it is important to demonstrate in the first step that, for the chosen experimental setup, retrieval practice indeed reduces unpracticed items' recognition. Therefore, in an initial experiment, we re-examined the effects of retrieval practice on recall and recognition of related unpracticed items to provide baseline results for the subsequent analysis of the effects of restudy when supplemented with context change.

Fig. 1 gives an overview of the three experiments. The three experiments were identical with respect to the initial study phase and the final test phase, but they differed in the intermediate and practice phases. In all three experiments, participants studied a categorized item list. An intermediate phase followed, which included backward counting and calculations (Experiment 1, see Fig. 1a) or context change tasks, like imagination (Experiment 2, see Fig. 1b) and semantic generation (Experiment 3, see Fig. 1c). In the practice phase, a subset of the studied items was practiced by either retrieval (Experiment 1) or restudy trials (Experiments 2 and 3). In the final test phase, participants' memory for all the items of the study list was tested using cued recall or item recognition testing (see Fig. 1d).

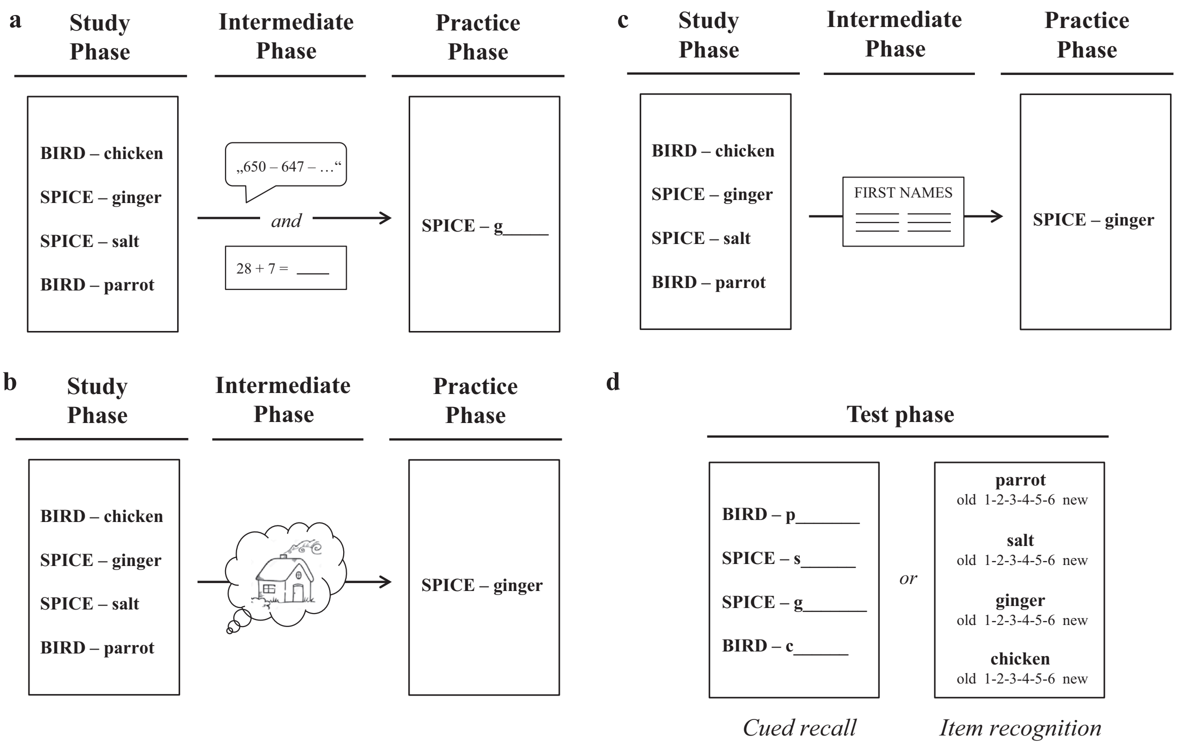

Fig. 1. Study, intermediate, practice, and test phases of the three experiments. (a) Study, intermediate, and practice phases of Experiment 1. Participants studied a categorized item list, engaged in counting and calculations, and then retrieved a subset of the items given the category label and the item's initial letter as retrieval cues. (b) Study, intermediate, and practice phases of Experiment 2. Participants studied a categorized item list, performed mental imagination tasks (e.g., being in the parents' house), and then restudied a subset of the items together with their category label. (c) Study, intermediate, and practice phases of Experiment 3. Participants studied a categorized item list, performed semantic generation tasks (generating exemplars of a given nonstudied category, e.g., first names), and then restudied a subset of the items together with their category label. (d) Test phase of all three experiments. Participants completed either a cued recall test, in which the category label and the item's initial letter were provided as retrieval cues, or a recognition test, in which they rated on a 6-point scale whether the item had been studied (old) or not (new).

Across the three experiments, we used the same materials with the same practiced, unpracticed, and control items as well as the same lure items, and the same design varying item type and test type within participants. Thus, if the effects of practice on unpracticed items varied across the three experiments, the variation should have been caused by the differences in the intermediate and practice phases.

We expected to replicate previous RIF studies in Experiment 1, observing RIF in both recall and item recognition (e.g., Anderson et al., 1994; Hicks & Starns, 2004). If so, on the basis of the context account of RIF and the implied equivalence hypothesis, restudy preceded by context change should also induce forgetting of the unpracticed items. Thus, RIF-like forgetting should arise with the imagination task (Experiment 2) as well as the semantic generation task (Experiment 3) in both recall and item recognition. Alternatively, if RIF and RIF-like forgetting arose in recall, and RIF, but not RIF-like forgetting, arose in item recognition, this would challenge the equivalence proposal of the context account, indicating that RIF in item recognition is not mediated by context change. Rather, such result would be in line with the retrieval specificity proposal of the inhibition account and suggest a critical role of retrieval and inhibition in RIF.

Experiment 1

Previous studies repeatedly demonstrated that retrieval practice on a subset of previously studied items can impair both recall and item recognition of related unpracticed items (see Murayama et al., 2014). The goal of Experiment 1 was to replicate this finding for the material and procedure employed in the present study. Using the standard retrieval practice task (Anderson et al., 1994), participants studied category-exemplar pairs, engaged in retrieval practice of a subset of items, and after a retention interval were tested on all studied items, using either cued recall or item recognition testing. Critically, no context change task preceded the practice phase. We expected retrieval practice to improve both recall and recognition of the practiced items, and to impair both recall and recognition of the unpracticed items.

Method

Participants

Forty-eight students of Regensburg University took part in the experiment (*M* = 22.83 years, *range* = 18—29 years, 43 female). They spoke German as native language. Monetary reward was provided in exchange for participation.

Materials

We drew sixteen semantic categories with six to-be-studied items and six lure items from published German word norms (Mannhaupt, 1983; Scheithe & Bäuml, 1995). Categories were allocated to one of two item sets (set 1: STATES OF THE U.S.A., MUSICAL INSTRUMENTS, FLOWERS, INSECTS, CAR EQUIPMENT, FRUITS, BIRDS, SPICES; set 2: AFRICAN STATES, KINDS OF FISH, PROFESSIONS, HOBBIES, TREES, KITCHEN EQUIPMENT, FOUR-LEGGED ANIMALS, ARTICLES OF CLOTHING). Additionally, we selected three categories (set 1: PARTS OF GRAMMAR, SANITARY ARTICLES, TOYS; set 2: ALCOHOLIC BEVERAGES, PARTS OF THE BODY, RELATIVES) with two exemplars each used as buffer items in the study and recognition lists. The German translations of the category labels of the sixteen experimental categories consisted of a single word. The to-be-studied exemplars within each category had a unique initial letter. With respect to their frequency in the word norms, items were alternately assigned to be study items or to be lure items. The medians of the studied items were 12.5 (set 1) and 14.0 (set 2); the medians of the lure items were 14.5 (set 1) and 11.0 (set 2).[1]

Design

The experiment had a 3 × 2 design with the within-participant factors of ITEM TYPE (practiced, unpracticed, control) and TEST TYPE (cued recall, recognition). The experiment consisted of two blocks that were identical apart from materials (set 1 or set 2) and test type (cued recall or recognition). We counterbalanced the order of test type and material across participants. In both blocks, participants completed four main phases: an initial study phase, an intermediate phase, a practice phase, and a final test phase. In the practice phase, participants practiced three exemplars of four categories. The remaining four categories served as control categories. Thus, three types of items were created: practiced items (*p*+); unpracticed items of practiced categories, i.e., items that were members of the same category as the *p*+ items but were not retrieved in the practice phase (*p*−); and items from unpracticed categories that served as controls for the practiced (*c*+) and unpracticed (*c*−) items. We counterbalanced categories across participants to be either practiced or not practiced (control). Thus, items that were practiced (*p*+ items) by half of the participants served as unpracticed control items (*c*+ items) for the other half of the participants (analogously for *p*- and *c*− items).

For the final recognition test, four further item types were generated, i.e., exemplars of practiced categories (*p*+ and *p*− lures) and of control categories (*c*+ and *c*− lures) that had not been presented in any other phase of the experiment. As described in more detail below, the differentiation between *p*+ and *p*− lures and *c*+ and *c*− lures merely stems from testing position in the recognition test: *p*− and *c*− lures were presented in the first half of the recognition test alongside the *p*− and *c*− items; *p*+ and *c*+ lures were presented in the second half alongside the *p*+ and *c*+ items. In the respective test type condition, all items were tested within the same single recognition test.

Procedure

Participants completed two blocks, with a five-minute break between blocks. In the study phase of each block, participants studied category-exemplar pairs (e.g., BIRD - *chicken*, SPICE - *ginger*, SPICE - *salt*) at a 4 s rate (ISI = 500 ms) displayed on a computer screen. The order of word pairs in the study list was blocked randomized: We compiled six blocks, each block comprised one exemplar from each category. Order of blocks and order of word pairs within the blocks were random. Three buffer items were presented at the beginning and ending of the study list.

The study phase was followed by an intermediate phase: After studying the category-exemplar pairs, participants counted backwards in steps of three from a three-digit number for 60 s. Subsequently, we provided simple math tasks (addition and subtraction of two two-digit numbers) for further 3 min. Thus, the intermediate phase in each block took 4 min. These tasks were used to approximately match the time frame of the context change tasks employed in Experiments 2 and 3, and were supposed to not induce any context change (e.g., Klein, Shiffrin, & Criss, 2007). In the subsequent practice phase, half of the exemplars from half of the categories were retrieved from memory: The category label and the initial letter of an exemplar were presented for 4 s (ISI = 500 ms; e.g., SPICE - *g*_) and participants were instructed to recall the matching exemplar from the study list orally while the experimenter logged the data. Presentation was blocked randomized. The twelve exemplars were practiced twice in consecutive cycles. No feedback was provided. Before the final test, participants worked on a distractor task for another 4 min (Frankfurter Aufmerksamkeitsinventar 2, FAIR-2; Moosbrugger, Oehlschlägel, & Steinwascher, 2011).

At test, participants engaged either in a cued recall or an item recognition test. In the cued recall test, the studied items were cued with the category label and the initial letter of the exemplar (e.g., SPICE - *s*_). Participants were asked to orally respond with the corresponding item within 5 s (ISI = 500 ms). The experimenter recorded the answers. Unpracticed items of practiced categories (*p*−) and the corresponding control items (*c*−) were tested first in order to avoid confounding output interference effects from the practiced items. We arranged the test list compiling six blocks, three blocks with exclusively *p*− and *c*− items and the other three blocks with the practiced items (*p*+) and their counterparts (*c*+). For each block, one exemplar from each category was drawn randomly. Order of the three first and the three last blocks as well as of the items within each block was random. At the beginning of the test, three of the six buffer items were tested in order to familiarize participants to the procedure.

The recognition test followed the procedure employed in Rupprecht and Bäuml (2016). All exemplars from the study

[1] English translations of the materials (originally in German) are available on request via e-mail to the authors.

list interspersed with lures were presented. Underneath each item, in the lower third of the screen, a schematic rating scale was displayed. Participants rated their confidence of an item having been previously studied (old) or not (new) on a 6-point scale (1 = *definitely old*, 6 = *definitely new*) (for arguments in favor of this rating procedure compared to a procedure that requires participants to make binary old/new decisions only, see Macmillan & Creelman, 2004; Parks & Yonelinas, 2008). Responses were typed in by the participants at their own pace, i.e., the next item did not appear on the screen until the participant had rated the presently displayed exemplar. Following Jonker et al. (2013), this was done to enhance participants' recollection and usage of context information during test. Data were logged automatically by the computer. Order of the recognition list was blocked randomized with two restrictions: old and new items were presented at most three times in a row; the first half of the list contained unpracticed items of practiced categories ($p-$), their control counterparts ($c-$), and corresponding lures, again in order to avoid confounding output interference effects from the practiced items. We compiled twelve blocks: six blocks consisting of $p-$ items, $c-$ items, $p-$ lures, and $c-$ lures constituting the first half of the test; and six blocks containing $p+$ items, $c+$ items, $p+$ lures, and $c+$ lures that were presented in the second part of the recognition test. For each block, one exemplar of each category was drawn and arrayed pseudo-randomly considering the above-mentioned restrictions. The six blocks within one test half were randomly drawn. At the beginning of the recognition list, three buffer items were presented.

Statistical analysis

For the cued recall test, correct recall and intrusion rates were analyzed with respect to differences in means between $p+$ items and $c+$ items to account for the beneficial effects of practice, and between $p-$ items and $c-$ items to account for the detrimental effects of practice.

For the recognition test, proportion of correctly recognized target items (i.e., hit rate) and proportion of incorrectly recognized lure items (i.e., the false alarm rate) were accumulated across the rating scale starting at the most confident criterion, i.e., *definitely old* ("1"). This procedure leads to an empirical Receiver Operating Characteristic (ROC) curve that relates hit rates and false alarm rates across variations in response criteria (i.e., the propensity to make a positive recognition response; e.g., Macmillan & Creelman, 2004; Parks & Yonelinas, 2008). With the present 6-point scale, hit and false alarm rates under five different response criteria arose. The first point of the ROC ("1") shows hit and false alarm rates when adopting the strictest response criterion, and each subsequent point ("2", "3", "4", "5") reflects performance at a more and more relaxed response criterion. Importantly, the function is cumulative and so both hit and false alarm rates are constrained to increase or remain constant as the scoring criterion is relaxed.

Analysis of recognition data followed Rupprecht and Bäuml (2016). In the first step, recognition data were analyzed separately for the single response criteria. Corrected hits (hits - false alarms) were calculated for each combination of item type and criterion, and, using ANOVA, it was examined for the three most conservative ("old") response criteria ("1", "2", "3") whether corrected hits for practiced items ($p+$) exceeded corrected hits for the corresponding control items ($c+$), and whether corrected hits for unpracticed items from practiced categories ($p-$) were lower than for the respective controls ($c-$). Analysis of corrected hits implicitly assumes that the ROC function is linear (e.g., Wixted, 2007b). However, because ROC functions are typically curvilinear and asymmetric, as is also the case in the present study (see below), analysis of corrected hits can serve as a first approximation only towards analysis of participants' recognition performance.

In the second step, recognition data were therefore analyzed using a signal detection approach, which takes the curvilinear and asymmetric form of the ROC into account. We presumed unequal variance for the distribution of old and new items to account for the typically asymmetrical shape of the ROC and, thus, we described the data by applying the unequal-variance signal detection model (e.g., Dunn, 2004; Wixted, 2007a). This model bases recognition judgments upon a single source of information, i.e., the items' general memory strength, which does not necessarily imply a single underlying memory process but, for instance, may reflect the additive or nonadditive combination of familiarity and recollection codes (e.g., Kelley & Wixted, 2001; Wixted & Stretch, 2004). Whenever an item exceeds the response criterion c_i, which is related to a particular level of confidence i, but does not exceed criterion c_{i-1}, participants rate the item accordingly with i. The memory strength of old items relative to new items can be derived from the distance between the means of the underlying strength distributions of those old and the new items (d_a). Applying the model to our 5-point ROC data, it results in seven free parameters (memory strength of old items d_a, variance of the distribution of old items σ, and five criterion points c_1-c_5) and consequently three degrees of freedom when testing the model's goodness of fit. Consistent with the prior work (e.g., Rupprecht & Bäuml, 2016; Spitzer & Bäuml, 2007), we applied the model to the group data for estimating the model parameters and adopted maximum-likelihood methods that could be used for statistical testing as well.

We firstly analyzed whether the unequal-variance signal detection model described the data sufficiently for each item type and practice condition. Then, in order to reveal potential beneficial and detrimental effects of practice, it was tested whether memory strength d_a varied significantly across item types and practice conditions. In particular, we examined whether d_a for practiced items ($p+$) exceeded d_a for the corresponding control items ($c+$), and whether d_a for unpracticed items from practiced categories ($p-$) was lower than for the respective controls ($c-$). We also examined whether the model's other parameters varied across item type.

Results

Practice phase

During retrieval practice, participants successfully retrieved 63.4% ($SD = 0.15$) of the items on the first practice cycle and 64.7% ($SD = 0.15$) of the items in total.

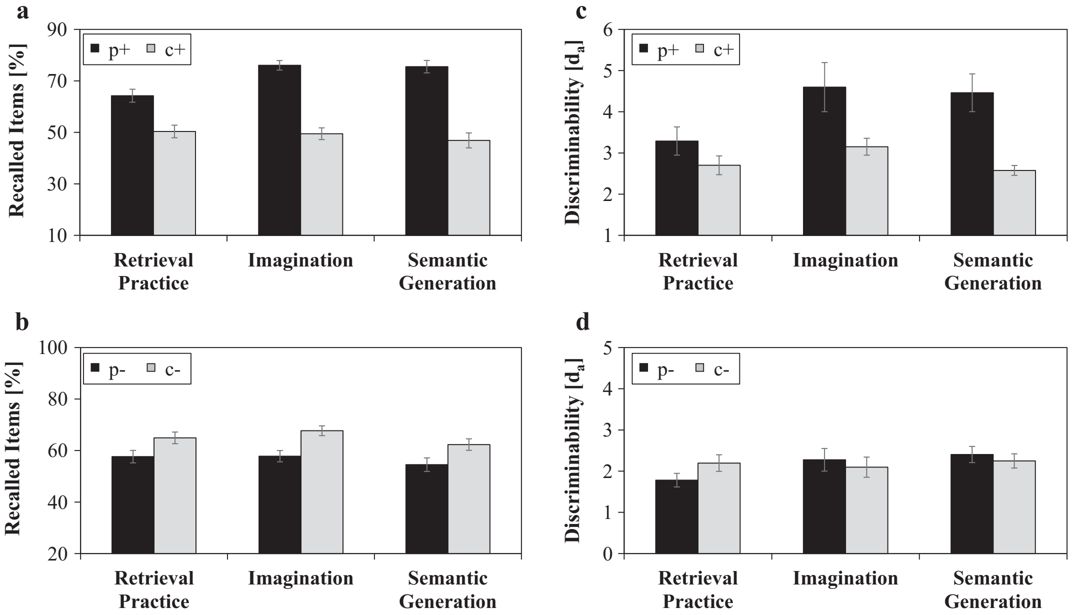

Fig. 2. Recall percentages and discriminability d_a for practiced items, unpracticed items, and control items. (a) Recall percentages for retrieved items (*p+*) and control items (*c+*) of Experiment 1 (left panel), restudied items (*p+*) and control items (*c+*) of Experiment 2 (middle panel), and restudied items (*p+*) and control items (*c+*) of Experiment 3 (right panel). (b) Recall percentages for unpracticed items from retrieval practiced categories (*p*−) and control items (*c*−) of Experiment 1 (left panel), unpracticed items from restudied categories (*p*−) and control items (*c*−) of Experiment 2 (middle panel), and unpracticed items from restudied categories (*p*−) and control items (*c*−) of Experiment 3 (right panel). (c) Discriminability d_a for retrieved items (*p+*) and control items (*c+*) of Experiment 1 (left panel), restudied items (*p+*) and control items (*c+*) of Experiment 2 (middle panel), and restudied items (*p+*) and control items (*c+*) of Experiment 3 (right panel). (d) Discriminability d_a for unpracticed items from retrieval practiced categories (*p*−) and control items (*c*−) of Experiment 1 (left panel), unpracticed items from restudied categories (*p*−) and control items (*c*−) of Experiment 2 (middle panel), and unpracticed items from restudied categories (*p*−) and control items (*c*−) of Experiment 3 (right panel). Error bars represent standard errors. For d_as, standard errors were approximated using bootstrapping with 5000 samples.

Recall test

Fig. 2a and b (left panels) show percentages of correctly recalled practiced (*p+*) and unpracticed (*p*−) items, together with their corresponding control (*c+*, *c*−) items. Regarding the beneficial effect of practice, participants, on average, recalled 64.2% of the *p+* items and 50.3% of the *c+* items. Recall levels differed significantly, $t(47) = 4.794$, $p < .001$, $d = 0.808$, indicating that practice was successful. Intrusion rates were .07 ($SD = 0.07$) for the *p+* items and .06 ($SD = 0.08$) for the *c+* items, and were not significantly different, $t(47) < 1$.

Regarding the detrimental effect of practice, on average, 57.6% of the *p*− items and 64.9% of the *c*− items were recalled. Recall performance was reliably reduced for the *p*− items, $t(47) = 3.157$, $p = .003$, $d = 0.450$, indicating the presence of RIF. Intrusion rates were .07 ($SD = 0.07$) for the *p*− items and .08 ($SD = 0.08$) for the *c*− items and did not differ reliably, $t(47) < 1$.

Recognition test: ANOVA of corrected hits

In Table 1, mean hit rates, false alarm rates, and corrected hit rates are displayed as a function of the five response criteria and the four item types. In the first step, we conducted ANOVAs to analyze for the three most conservative ("old") response criteria whether corrected hits varied with item type. Regarding the effects of retrieval practice on the practiced (*p+*) items relative to their controls (*c+*), a 2 × 3 ANOVA with the within-participants factors of ITEM TYPE (*p+*, *c+*) and RESPONSE CRITERION ("1", "2", "3") showed a main effect of ITEM TYPE, $F(1,47) = 17.723$, $MSE = 0.056$, $p < .001$, $\eta^2 = 0.274$, with higher corrected hits for the practiced than the control items, indicating that practice was successful. The effect did not vary with response criterion, $F(2,94) = 1.726$, $MSE = 0.003$, $p = .184$, $\eta^2 = 0.035$.[2] In contrast to corrected hits, false alarm rates did not depend on item type. A 2 × 3 ANOVA with the factors of ITEM TYPE (*p+*, *c+*) and RESPONSE CRITERION ("1", "2", "3") showed no main effect of ITEM TYPE, $F(1,47) < 1$, and no interaction between the two factors, $F(2,94) < 1$.

Regarding the effects of retrieval practice on corrected hits of the unpracticed (*p*−) items relative to corrected hits of their controls (*c*−), a 2 × 3 ANOVA with the factors of ITEM TYPE (*p*−, *c*−) and RESPONSE CRITERION ("1", "2", "3") showed a main effect of ITEM TYPE, $F(1,47) = 12.984$, $MSE = 0.045$, $p = .001$, $\eta^2 = 0.216$, with lower corrected hits for *p*− items, but no interaction between the two factors, $F(2,94) = 1.498$, $MSE = 0.008$, $p = .229$, $\eta^2 = 0.031$. These results indicate that retrieval practice induced RIF in item

[2] In this ANOVA and all forthcoming related ANOVAs of Experiments 1–3, there was also a main effect of RESPONSE CRITERION. However, because this effect is kind of trivial since ROC functions are cumulative (see section 'Method' above), we do not report detailed results on it.

Table 1
Hit rates, false alarm rates, and corrected hit rates for Experiment 1.

Item type		Response criteria				
		"1"	"2"	"3"	"4"	"5"
p+	Hits	.797	.866	.903	.945	.974
	False alarms	.035	.057	.130	.252	.488
	Corrected hits	.762	.809	.773	.693	.486
c+	Hits	.672	.741	.797	.865	.922
	False alarms	.029	.064	.125	.257	.457
	Corrected hits	.643	.677	.672	.608	.465
p−	Hits	.602	.719	.783	.859	.929
	False alarms	.087	.151	.238	.377	.616
	Corrected hits	.515	.568	.545	.482	.313
c−	Hits	.712	.800	.851	.899	.958
	False alarms	.082	.148	.234	.345	.592
	Corrected hits	.630	.652	.617	.554	.366

Notes. (Corrected) hit and false alarm rates are shown as a function of item type and response criterion. p+ = practiced items; c+ = unpracticed items from unpracticed categories; p− = unpracticed items from practiced categories; c− = unpracticed items from unpracticed categories. "1" reflects the strictest response criterion, i.e., definitely old, and each subsequent number ("2", "3", etc.) reflects a more and more relaxed criterion. Corrected hits = hits − false alarms.

recognition. False alarm rates did not vary with item type, as is indicated by the results of a 2 × 3 ANOVA with the factors of ITEM TYPE ($p-$, $c-$) and RESPONSE CRITERION ("1", "2", "3"), which showed no main effect of ITEM TYPE, $F(1,47) < 1$, and no interaction between the two factors, $F(2,94) < 1$.

Recognition test: analysis of hit and false alarm rates using the unequal-variance signal detection model

In the second step, we employed the unequal-variance signal detection model to analyze the data, which takes the curvilinear and asymmetric form of the ROC into account. The model fit the recognition data of the four types of items well, all $\chi^2s(3) < 1.408$, all $ps > .703$. Fig. 2c and d (left panels) show estimated discriminability d_a for practiced (p+) and unpracticed ($p-$) items, together with d_a for the corresponding control (c+, $c-$) items. Retrieval practiced (p+) items showed enhanced d_a relative to the control (c+) items, $\chi^2(1) = 3.904$, $p = .048$, indicating that practice was successful. Relative to the $c-$ items, retrieval practice reduced d_a for the unpracticed ($p-$) items, $\chi^2(1) = 5.374$, $p = .020$, indicating RIF.

For both the practiced items and their controls, and the unpracticed items and their controls, the variance of the old items' distribution, as estimated by parameter σ, did not vary significantly across item type, both $\chi^2s(1) < 0.504$, $ps > .477$, but was larger than 1.0, both $\chi^2s(1) > 27.905$, $ps < .001$, indicating that the model's assumption of unequal variances for old and new items improved the description of the data significantly. For both pairs of item types, the placement of the five confidence criteria did not vary with item type, $\chi^2s(5) < 2.244$, $ps > .814$.[3]

Discussion

Retrieval practice improved memory for the practiced items and impaired memory for the unpracticed items, both when using recall and when using item recognition as memory tests. The findings replicate prior work, which repeatedly demonstrated the benefits and costs of selective retrieval practice (e.g., Anderson et al., 1994; Hicks & Starns, 2004; see also Murayama et al., 2014). The results are consistent with the theoretical view that inhibition operates during practice and reduces unpracticed items' memory strength, so that memory of these items is impaired over a wide range of memory tests. But the results are also compatible with the alternative view that retrieval practice prompts a context change and this change underlies the memory impairment for the unpracticed items. While Experiment 1 thus must leave it open whether inhibition or context change mediated the induced forgetting, it can serve as an important baseline for Experiments 2 and 3, indicating that, in the present experimental setup, putative context effects should become visible in both recall and item recognition. Experiments 2 and 3 examined the effects of restudy when preceded by context change on recall and recognition of unpracticed items.

Experiment 2

Jonker et al. (2013) examined the effects of restudy preceded by context change on recall of the unpracticed items employing an imagination task to induce context shift. In their experiment, participants studied a list of categorized items. Then the imagination task followed and participants were asked to imagine walking through their parents' house, before they restudied a subset of the items. Finally, all items were tested employing a cued recall test. Results showed that restudy when preceded by the imagination task produced RIF-like forgetting and impaired recall of the unpracticed items, thus mimicking the typical effect of retrieval practice on unpracticed items. The goal of

[3] Half of the participants in this experiment started testing with the cued recall test and the other half with item recognition. When restricting analyses to the data of participants' first memory test, exactly the same pattern of results arose as reported above, indicating that testing order did not influence the results. The same held true for the results of Experiments 2 and 3 below.

Experiment 2 was to replicate this finding for recall and examine whether it generalizes to item recognition. Like Jonker et al. (2013), we employed imagination tasks to induce context change.

On the basis of the results of Experiment 1 and the proposal incorporated in the context account that restudy when supplemented by preceding context change can simulate the effects of retrieval practice, we expected that restudy preceded by context change impaired both recall and item recognition of the unpracticed items. Alternatively, if restudy preceded by context change induced forgetting in recall but not in item recognition, this would indicate that context change did not mediate the effects of retrieval practice in Experiment 1 and that, in general, retrieval practice and restudy preceded by context change trigger different processes and induce different effects on the unpracticed items.

Method

Participants

Further 48 students of Regensburg University participated in the experiment ($M = 21.44$ years, *range* $= 18-29$ years, 37 female). All participants spoke German as native language and received money in exchange for participation.

Materials

The same material as in Experiment 1 was employed.

Design

The experiment had the same 3 × 2 design as Experiment 1 with the within-participants factors of ITEM TYPE (practiced, unpracticed, control) and TEST TYPE (cued recall, recognition). The only differences between the two experiments were the nature of the intermediate task and type of practice: In the intermediate phase, participants were engaged in two successive imagination trials rather than participating in counting and calculation tasks, and in the practice phase they restudied a subset of the studied items rather than retrieving these items (see Fig. 1b). Analogous to Experiment 1, restudied items are denoted *p+* items and unpracticed items of restudied categories are denoted *p−* items. Respective control items are again denoted *c+* and *c−* items. Lures belonging to restudied categories are denoted *p+ lures* and *p− lures*.

Procedure

Study phase, distractor task, and test phase did not differ from Experiment 1. In the intermediate phase, a context change task was administered. Participants were instructed to imagine a scenario as vividly as possible and to write it down within 2 min. Four imagination tasks were employed with two tasks in each experimental block (being in the parents' house; recalling a happy childhood event; winning 10 million Euro in the lottery; being able to perform magic; see Delaney, Sahakyan, Kelley, & Zimmerman, 2010; Sahakyan & Kelley, 2002). The tasks were randomly assigned across the two tests. Participants completed a block's two imagination tasks consecutively. Following this intermediate phase, participants practiced half of the exemplars from half of the categories by extra study. The complete category-exemplar pair was reexposed on the computer screen for 4 s (ISI = 500 ms; e.g., SPICE - *ginger*). Like in Jonker et al. (2013), participants were asked to read the pairs out loud and to restudy them as thoroughly as possible. Order of presentation was blocked randomized. The twelve pairs were practiced in two consecutive cycles. The final test, cued recall or item recognition, followed after the same 4 min distractor task as was used in Experiment 1.

Statistical analysis

Statistical data analysis was analogous to Experiment 1.

Results

Recall test

Percentages of correctly recalled practiced (*p+*) and unpracticed (*p−*) items, together with their corresponding control (*c+*, *c−*) items, are displayed in Fig. 2a and b (middle panels). Regarding the beneficial effect of practice, mean recall rates mounted up to 76.0% and 49.5% for the *p+* and *c+* items. The numerical difference was reliable, $t(47) = 11.434$, $p < .001$, $d = 1.847$, suggesting that practice improved recall. Intrusion rates were .04 ($SD = 0.06$) for the *p+* items and .07 ($SD = 0.10$) for the *c+* items, and did not differ significantly, $t(47) = 1.785$, $p = .081$.

Regarding the detrimental effect of practice, participants recalled 57.8% of the *p−* items and 67.7% of the *c−* items, $t(47) = 4.189$, $p < .001$, $d = 0.696$, showing significant recall impairment for the *p−* items and thus RIF-like forgetting. Intrusion rates were .07 ($SD = 0.09$) for the *p−* items and .09 ($SD = 0.09$) for the *c−* items and did not vary significantly, $t(47) < 1$.

Recognition test: ANOVA of corrected hits

Table 2 depicts mean hit rates, false alarm rates, and mean corrected hit rates as a function of response criterion and item type. Regarding the beneficial effects of practice on corrected hits, a 2 × 3 ANOVA with the within-participants factors of ITEM TYPE (*p+*, *c+*) and RESPONSE CRITERION ("1", "2", "3") revealed a main effect of ITEM TYPE, $F(1, 47) = 36.786$, $MSE = 0.037$, $p < .001$, $\eta^2 = 0.439$, which was qualified by an interaction with the factor of response criterion, $F(2, 94) = 7.600$, $MSE = 0.004$, $p = .001$, $\eta^2 = 0.139$. However, corrected hit rates for *p+* items exceeded corrected hit rates for *c+* items for all three response criteria, all $ts(47) > 4.252$, all $ps < .001$, all $ds > .445$, indicating that practice was successful. False alarm rates did not depend on item type. A 2 × 3 ANOVA with the factors of ITEM TYPE (*p+*, *c+*) and RESPONSE CRITERION ("1", "2", "3") showed no main effect of ITEM TYPE,

Table 2
Hit rates, false alarm rates, and corrected hit rates for Experiment 2.

Item type		Response criteria				
		"1"	"2"	"3"	"4"	"5"
p+	Hits	.957	.971	.979	.986	.995
	False alarms	.063	.099	.168	.261	.467
	Corrected hits	.894	.872	.811	.726	.528
c+	Hits	.769	.818	.858	.894	.938
	False alarms	.049	.082	.151	.257	.457
	Corrected hits	.721	.736	.707	.637	.481
p−	Hits	.701	.769	.819	.858	.917
	False alarms	.094	.134	.207	.321	.521
	Corrected hits	.608	.635	.613	.537	.396
c−	Hits	.719	.781	.844	.891	.943
	False alarms	.090	.137	.198	.293	.479
	Corrected hits	.628	.644	.646	.597	.464

Notes. (Corrected) hit and false alarm rates are shown as a function of item type and response criterion. p+ = practiced items; c+ = unpracticed items from unpracticed categories; p− = unpracticed items from practiced categories; c− = unpracticed items from unpracticed categories. "1" reflects the strictest response criterion, i.e., definitely old, and each subsequent number ("2", "3", etc.) reflects a more and more relaxed criterion. Corrected hits = hits – false alarms.

$F(1,47) = 1.201$, $MSE = 0.016$, $p = .279$, $\eta^2 = 0.025$, and no interaction between the two factors, $F(2,94) < 1$.

Regarding the detrimental effects of practice on corrected hits, a 2×3 ANOVA with the factors of item type ($p-$, $c-$) and response criterion ("1", "2", "3") showed no main effect of item type, $F(1,47) < 1$, and no interaction between the two factors, $F(2,94) < 1$, indicating that restudy preceded by context change did not impair recognition of the unpracticed items. False alarm rates did also not depend on item type, as is indicated by the results of a 2×3 ANOVA with the factors of item type ($p-$, $c-$) and response criterion ("1", "2", "3"), which showed no main effect of item type, $F(1,47) < 1$, and no interaction between the two factors, $F(2,94) < 1$.

Next to null-hypothesis significance testing, we addressed the issue of detrimental effects of practice on corrected hits by estimating the Bayes factor, reflecting the odds in favor of the null hypothesis, and the conditional probability of the null hypothesis given the present corrected hits data for the unpracticed and corresponding control items. We found a Bayes factor of $BF = 5.483$ in favor of the null hypothesis and a conditional probability of the null hypothesis given the data of $p(H_0|D) = .846$. The Bayes factor suggests that the data were 5.5 times more likely to occur under the null hypothesis than under the alternative hypothesis that assumes differences in corrected hits between unpracticed and control items. According to Raferty (1995), the conditional probability indicates evidence of medium size in favor of the null hypothesis.

Recognition test: analysis of hit and false alarm rates using the unequal-variance signal detection model

The unequal-variance signal detection model described the data of the four item types well, all $\chi^2 s(3) < 1.086$, all $ps > .780$. Fig. 2c and d (middle panels) show discriminability d_a for practiced ($p+$), unpracticed ($p-$), and control ($c+$, $c-$) items. Practiced ($p+$) items showed higher d_a than the control ($c+$) items, $\chi^2(1) = 4.415$, $p = .036$, indicating improved recognition of the practiced items after the restudy trials. Critically, however, restudy preceded by imagination did not impair d_a for unpracticed ($p-$) items relative to the control ($c-$) items, $\chi^2(1) = 0.601$, $p = .438$, again indicating that no RIF-like forgetting arose in item recognition.

We did not find significant differences in σ between $p+$ and $c+$ items and between $p-$ and $c-$ items, $\chi^2 s(1) < 2.827$, $ps > .130$. Like in Experiment 1, σ was significantly larger than 1.0, $\chi^2 s(1) > 25.398$, $ps < .001$, indicating that the model's assumption of unequal variance for old and new items improved the description of the data significantly. Differences in the placement of the five confidence criteria between $p+$ and $c+$ items and between $p-$ and $c-$ items, did not reach significance, $\chi^2 s(5) < 3.041$, $ps > .693$.

Discussion

The results show that imagination prior to restudy can improve recall of the practiced items and reduce recall of the unpracticed items, thus replicating the results reported in Jonker et al. (2013). Extending this prior work, the results, however, also reveal that the forgetting of the unpracticed items does not generalize to item recognition. Indeed, while restudy preceded by imagination improved recognition of the practiced items, it left recognition of the unpracticed items unaffected. On the basis of the results of Experiment 1, this finding suggests that restudy preceded by imagination can have the same effects on recall of the practiced and unpracticed items as retrieval practice has. However, with respect to item recognition, retrieval practice but not restudy preceded by imagination induces forgetting of the unpracticed items. This finding provides evidence for an experimental dissociation between the two practice tasks and indicates that context change did not mediate the RIF effect in Experiment 1. However, before drawing more firm conclusions on the role of context change in RIF, we aimed to replicate the findings of Experiment 2 using a different method than imagination to induce context change.

Experiment 3

Imagination tasks have repeatedly been used in the literature to induce context change (Delaney et al., 2010; Jonker et al., 2013; Pastötter & Bäuml, 2007; Sahakyan & Kelley, 2002), but there are alternative ways to promote such change. For instance, several studies employed semantic generation tasks to change participants' internal context. Divis and Benjamin (2014), for instance, used a multiple-list learning paradigm to study how semantic generation of unstudied category exemplars between study of the single lists affects recall of the first and last study lists. Semantic generation between lists led to higher recall rates for the final list that followed the generation cycles and to lower recall rates for the first list that preceded the cycles, which is consistent with the view that, like imagination, semantic generation can induce context change (for related results, see Jang & Huber, 2008; Pastötter et al., 2011). The goal of Experiment 3 was therefore to replicate the results of Experiment 2 using semantic generation rather than imagination to change participants' internal context and to examine how subsequent selective restudy affects memory for the unpracticed items. We expected that restudy preceded by semantic generation would improve both recall and recognition of the practiced items. In particular, we expected that restudy would reduce recall, but would not reduce item recognition of the unpracticed items.

Method

Participants

Another 48 students were recruited at Regensburg University to participate in the experiment ($M = 23.25$ years, *range* $= 20$—29 years, 39 female). All spoke German as native language. Participation was rewarded monetarily.

Materials

Materials were identical to Experiments 1 and 2.

Design

A 3 × 2 design was employed varying ITEM TYPE (practiced, unpracticed, control) and TEST TYPE (cued recall, recognition) within participants. The only difference between the present experiment and Experiment 2 was that, in the intermediate phase, participants were asked to generate as many exemplars from semantic categories as possible rather than engaging in an imagination task. In the practice phase, participants again restudied a subset of the studied items. Analogous to Experiment 2, *p+* items represent restudied items, *p−* items represent unpracticed items of restudied categories, and *c+* and *c−* items represent corresponding control items. Again, *p+ lures* and *p− lures* label foils that are members of restudied categories.

Procedure

Study phase, practice phase, distractor task, and test phase were identical to Experiment 2. However, the intermediate phase consisted of semantic generation tasks. Participants were instructed to think of as many exemplars from a particular category (colors, candy, first names, means of transport) as possible and write them down within 2 min. None of the to-be-studied items or lures belonged to one of the four semantic categories. Participants completed two out of the four semantic retrieval tasks consecutively, the remaining two tasks were presented in the second block of the experiment. Like in Experiment 2, participants then engaged in extra study of a subset of the word pairs. The complete category-exemplar pair was reexposed on the computer screen for 4 s (ISI = 500 ms; e.g., SPICE - *ginger*). Again, we asked participants to read the pairs out loud and to restudy them for a later test. The twelve pairs were practiced in two consecutive cycles in blocked randomized order. The final test, cued recall or item recognition, followed after the same 4 min distractor task as was used in Experiments 1 and 2.

Statistical analysis

The same statistical analyses as in Experiments 1 and 2 were employed.

Results

Recall test

Fig. 2a and b (right panels) show percentages of correctly recalled practiced (*p+*) and unpracticed (*p−*) items together with their corresponding control (*c+*, *c−*) items. Regarding the beneficial effect of practice, participants recalled on average 75.5% of the *p+* items and 46.9% of the *c+* items. Recall differed significantly between item types, $t(47) = 8.656$, $p < .001$, $d = 1.547$, indicating that practice was successful. Intrusion rates were .04 ($SD = 0.06$) for the *p+* items and .06 ($SD = 0.07$) for the *c+* items, and did not vary significantly between item types, $t(47) = 1.295$, $p = .202$.

Regarding the detrimental effect of practice, recall rates for *p−* items and *c−* items reached 54.5% and 62.3%, respectively. The numerical difference was reliable, $t(47) = 3.986$, $p < .001$, $d = 0.462$, suggesting that selective restudy preceded by semantic generation induced RIF-like forgetting. Intrusion rates were .06 ($SD = 0.08$) for the *p−* items and .08 ($SD = 0.07$) for the *c−* items, but the difference was not significant, $t(47) = 1.400$, $p = .168$.

Recognition test: ANOVA of corrected hits

Table 3 shows mean hit rates, false alarm rates, and mean corrected hit rates for the five response criteria and the four item types. Regarding the beneficial effect of restudy supplemented with prior semantic generation on corrected hits, we conducted a 2 × 3 ANOVA with the within-participants factors of ITEM TYPE (*p+*, *c+*) and RESPONSE

Table 3
Hit rates, false alarm rates, and corrected hit rates for Experiment 3.

Item type		Response criteria				
		"1"	"2"	"3"	"4"	"5"
p+	Hits	.953	.977	.988	.995	.995
	False alarms	.040	.069	.109	.214	.455
	Corrected hits	.913	.908	.879	.781	.540
c+	Hits	.700	.762	.814	.873	.946
	False alarms	.042	.061	.099	.184	.439
	Corrected hits	.658	.701	.715	.689	.507
p−	Hits	.663	.750	.804	.866	.941
	False alarms	.042	.099	.163	.307	.550
	Corrected hits	.622	.651	.641	.559	.391
c−	Hits	.684	.771	.821	.873	.946
	False alarms	.071	.116	.187	.335	.563
	Corrected hits	.613	.655	.634	.538	.384

Note. (Corrected) hit and false alarm rates are shown as a function of item type and response criterion. p+ = practiced items; c+ = unpracticed items from unpracticed categories; p− = unpracticed items from practiced categories; c− = unpracticed items from unpracticed categories. "1" reflects the strictest response criterion, i.e., definitely old, and each subsequent number ("2", "3", etc.) reflects a more and more relaxed criterion. Corrected hits = hits − false alarms.

CRITERION ("1", "2", "3"). We observed a main effect of ITEM TYPE, $F(1,47) = 71.723$, $MSE = 0.044$, $p < .001$, $\eta^2 = 0.604$, suggesting that practice was successful, and an interaction of the two factors, $F(2,94) = 9.097$, $MSE = 0.006$, $p < .001$, $\eta^2 = 0.162$. Although the size of the beneficial effect thus varied with the particular response criterion, the effect was present for each single criterion, all $ts(47) > 6.158$, all $ps < .001$, all $ds > 1.079$. In contrast to corrected hits, false alarm rates did not depend on item type. A 2 × 3 ANOVA with the factors of ITEM TYPE (*p+*, *c+*) and RESPONSE CRITERION ("1", "2", "3") showed no main effect of ITEM TYPE, $F(1,47) < 1$, and no interaction between the two factors, $F(2,94) < 1$.

Regarding the detrimental effect of restudy supplemented with prior semantic generation on corrected hits, a 2 × 3 ANOVA with the factors of ITEM TYPE (*p*−, *c*−) and RESPONSE CRITERION ("1", "2", "3") showed no main effect of ITEM TYPE, $F(1,47) < 1$, and no interaction between the two factors, $F(2,94) < 1$, indicating that restudy preceded by semantic generation did not affect recognition of the unpracticed items.[4] False alarm rates did also not vary with item type, as is indicated by the results of a 2 × 3 ANOVA with the factors of ITEM TYPE (*p*−, *c*−) and RESPONSE CRITERION ("1", "2", "3"), which showed no main effect of ITEM TYPE, $F(1,47) = 1.972$, $MSE = 0.021$, $p = .167$, $\eta^2 = 0.040$, and no interaction between the two factors, $F(2,94) < 1$.

Next to null-hypothesis significance testing, we again addressed the issue of detrimental effects of practice on corrected hits by estimating the Bayes factor, reflecting the odds in favor of the null hypothesis, and the conditional probability of the null hypothesis given the present corrected hits for the unpracticed and corresponding control items. We found an estimated Bayes factor of $BF = 6.847$ in favor of the null hypothesis and a conditional probability of the null hypothesis given the data of $p(H_0|D) = .873$. The Bayes factor reflects a 6.8 times higher likelihood for the present data to arise given the null hypothesis compared to the alternative hypothesis, and the conditional probability provides evidence of medium strength in favor of the null hypothesis.

Recognition test: analysis of hit and false alarm rates using the unequal-variance signal detection model

The unequal-variance signal detection model provided a good fit for the data of the four item types, all $\chi^2s(3) < 4.178$, all $ps > .242$. Fig. 2c and d (right panels) show discriminability d_a for practiced (*p+*), unpracticed (*p*−), and control (*c+*, *c*−) items. Regarding the beneficial effect of practice, d_a for the practiced (*p+*) items exceeded that of the control (*c+*) items, $\chi^2(1) = 16.828, p < .001$, suggesting that practice was successful. Regarding the detrimental effect of practice, d_a did not differ between *p*− items and *c*− items, $\chi^2(1) = 0.541$, $p = .462$, indicating that no RIF-like forgetting arose in recognition memory.

Like in Experiments 1 and 2, σ did not vary with item type, $\chi^2s(1) < 0.412$, $ps > .520$, and was larger than 1.0, $\chi^2s(1) > 25.531, ps < .001$. Again, the placement of the five confidence criteria did not differ between item types, $\chi^2s(5) < 7.593$, $ps > .180$.

Additional analyses

Experiments 1–3 differed in the intermediate and practice phases but were identical in all other aspects. We therefore directly compared results of the three experiments, in both recall and item recognition.

Recall tests

All three experiments above showed improved recall of practiced items, with the improvement effect being numer-

[4] Consistent with the finding that RIF-like forgetting is present in recall but is absent in item recognition, a 2 × 2 ANOVA with the factors of ITEM TYPE (*p*−, *c*−) and TEST TYPE (recall, item recognition) also showed a significant interaction between the two factors, $F(1,47) = 15.573$, $MSE = 0.690$, $p < .001$, $\eta^2 = 0.249$. The same interaction arose for the results of Experiment 2, $F(1,47) = 16.749$, $MSE = 0.975$, $p < .001$, $\eta^2 = 0.263$.

ically higher in Experiments 2 and 3 than in Experiment 1. Consistently, a 2 × 3 ANOVA with the within-participants factor of ITEM TYPE (*p*+, *c*+) and the between-participants factor of PRACTICE CONDITION (retrieval practice, restudy-plus-imagination, restudy-plus-semantic-generation) showed a significant interaction between the two factors, $F(2, 141) = 7.141$, $MSE = 0.019$, $p = .001$, $\eta^2 = .092$, indicating that the improvement effect differed also statistically. All three experiments above also showed impaired recall of unpracticed items, with the impairment effect being numerically similar across experiments. Results of a 2 × 3 ANOVA with the factors of ITEM TYPE ($p-$, $c-$) and PRACTICE CONDITION confirmed this finding, showing no significant interaction between the two factors, $F(2, 141) < 1$.

Recognition tests

All three experiments showed enhanced corrected hits for practiced items, with the enhancement effect being numerically higher in Experiments 2 and 3 than in Experiment 1. Consistently, 2 × 3 × 3 ANOVA with the within-participants factors of ITEM TYPE (*p*+, *c*+) and RESPONSE CRITERION ("1", "2", "3"), and the between-participants factor of PRACTICE CONDITION (retrieval practice, restudy-plus-imagination, restudy-plus-semantic-generation) showed a significant interaction between item type and practice condition, $F(2, 141) = 3.600$, $MSE = 0.046$, $p = .030$, $\eta^2 = .049$, suggesting that the improvement effect differed also statistically. In particular, Experiment 1 showed reduced corrected hit rates for unpracticed items, whereas Experiments 2 and 3 did not. The results of a 2 × 3 × 3 ANOVA with the factors of ITEM TYPE ($p-$, $c-$), RESPONSE CRITERION, and PRACTICE CONDITION confirmed this finding, showing a significant interaction between item type and practice condition, $F(2, 141) = 3.320$, $MSE = 0.052$, $p = .039$, $\eta^2 = .045$. Analysis of hit and false alarm rates using the unequal-variance signal-detection model led to the same conclusions, demonstrating that both the improvement effect for practiced items and the impairment effect for unpracticed items in discriminability d_a varied significantly across experiments, both $\chi^2 s(2) > 12.375$, both $ps < .002$.

Discussion

The results show that restudy with preceding semantic generation can improve recall of the practiced items, but reduce recall of the unpracticed items, thus generalizing the recall results reported in Experiment 2 from imagination to semantic generation. Moreover, like imagination, semantic generation improved recognition of the practiced items, but did not reduce recognition of the unpracticed items. Together, the results of Experiments 2 and 3 thus suggest that context change manipulations when incorporated in the retrieval practice paradigm create RIF-like forgetting in recall but not in item recognition. Because this finding contrasts with the effects of retrieval practice reported in Experiment 1, which showed RIF both in recall and item recognition, it indicates that RIF in Experiment 1 was not mediated by context change.

General discussion

The context account of RIF explains RIF in terms of a mismatch of study context and reinstated context at test for the unpracticed items, indicating that RIF is not retrieval specific and the effects of retrieval practice can be simulated by restudy when it is preceded by context change. Across three experiments, this study examined the adequacy of this proposal. The recall results of the present experiments are consistent with the context account. They show both retrieval practice and restudy when preceded by context change to enhance recall of the practiced items and to impair recall of the unpracticed items. This held when context change was induced by an imagination task (Experiment 2) and when it was induced by semantic generation (Experiment 3). In contrast, the recognition results of the present experiments disagree with the context account. Whereas selective retrieval was found to impair recognition of unpracticed items, selective restudy left recognition of the unpracticed items unaffected, both when it was preceded by an imagination task (Experiment 2) and when it was preceded by semantic generation (Experiment 3). This observation challenges the equivalence proposal incorporated into the context account, according to which restudy when supplemented with context change should mimic the effects of retrieval practice, and thus challenges the context account.

The present results are consistent with the assumption of a critical role of inhibition in RIF. The inhibition account claims that retrieval practice can create interference during practice and reduce interfering unpracticed items' memory representations in order to overcome the interference. Such reduction in memory strength is supposed to be retrieval specific, i.e., to arise in response to (competitive) retrieval practice but to not arise in response to restudy trials and context change, and should be visible over a wide range of memory tests, including recall and item recognition (e.g., Anderson, 2003). Although the present results show similar effects of retrieval and restudy in recall, RIF, but not RIF-like forgetting, was observed in item recognition, which supports the retrieval specificity proposal and strengthens the view of a critical role of retrieval and inhibition in RIF (see also below).

A conceivable reason for the divergent findings between the two memory tests employed in the present experiments could be that both retrieval practice and restudy when preceded by imagination or semantic generation induced context change, but that retrieval practice induced a larger amount of context change than did the imagination and semantic generation tasks. In such case, both RIF and RIF-like forgetting may arise in recall, but mainly RIF may be present in item recognition, at least if recall was more sensitive to context effects than item recognition. The results of the present recall tests make such proposal unlikely, however. Indeed, the amount of forgetting of the unpracticed items in the single recall tests did not vary significantly with type of practice (see section 'Additional analyses' above) and was numerically even larger after restudy (imagination: 9.9%; semantic generation: 7.8%) than after retrieval practice (7.7%). This indicates that the putative context change should have been at least as large after restudy than after retrieval practice. The difference in the effects of the single practice methods on recognition of the unpracticed items therefore cannot easily be attributed to reduced context change in the two restudy conditions.

Testing the unpracticed items prior to the practiced items, as was done in Jonker et al. (2013), the present study (see above), and most other RIF studies (see Murayama et al., 2014), may bias the results against the context account, because such procedure can increase the likelihood that, when recalling the unpracticed items, participants reinstate the study context - the only context that the unpracticed items were presented in -, rather than the practice context as the context account suggests (Jonker et al., 2013, Footnote 6). This holds particularly if the tests are not blocked by category, as was done in Jonker et al., but by item type, as was done in the present study. Our results nevertheless demonstrate both RIF and RIF-like forgetting in the recall test, with amount of RIF-like forgetting being not smaller than in Jonker et al.'s experiments (9.9% and 7.8% in the present experiments versus about 7% and 9% in Jonker et al.'s experiments), which indicates that the putative context change was successful.

The context account predicts the presence of RIF and RIF-like forgetting in item recognition if participants' recognition judgments were not speeded and participants were rather instructed to respond accurately (Jonker et al., 2013, p. 868). The recognition experiments in this study followed this premise and did not impose any temporal pressure on the participants' judgments. Doing so, a mean recognition latency for correctly recognized items (ratings "1", "2", or "3") of 1.721 ms arose. This latency fits with prior work by Verde and Perfect (2011), who reported hit recognition latencies in RIF using both self-paced and speeded recognition testing. While latencies in Verde and Perfect's speeded condition (587 ms) were nearly thrice as fast as in the present study, latencies in their self-paced condition (1.233 ms) were more similar to those observed in the present study, although they were still faster.[5] These numbers clearly indicate that recognition judgments were not speeded in the present study. The present results can thus provide a test of the context account and the observed discrepancy in the recognition findings between retrieval and restudy trials disagrees with the account.

Further challenges for the context account of RIF

Not only the present recognition findings but also the results of some previous recall studies challenge the context account. These studies investigated effects of variations in context between study and practice on RIF or RIF-like forgetting, thus coming up with some tests of the account. For instance, although originally not designed to test the context account, some studies examined whether RIF still arises when mental context change, as induced by negative moods or stress, precedes retrieval practice. Following the account's assumption that, during retrieval practice, participants' context is shifted away from the list, RIF should not disappear when (additional) context change precedes retrieval practice. If anything, amount of RIF may be expected to increase in such situation due to enhanced contextual drift. The results, however, demonstrated that RIF does no longer occur when context change precedes retrieval practice, which disagrees with the context account (e.g., Bäuml & Kuhbandner, 2007; Koessler, Engler, Riether, & Kissler, 2009; for related results, see Bäuml & Samenieh, 2012; Storm, Bjork, & Bjork, 2007).[6]

Soares, Polack, and Miller (2016) tested the context account employing other methods to vary the contextual overlap between study and retrieval practice. They either showed participants the to-be-retained items intact or participants generated the items during study, before they retrieval practiced some of the encoded items (Experiment 1). Alternatively, participants were provided the same or different salient item-specific features (font, color, etc.) in the study and practice phases (Experiment 2). The rationale was that if processes and task demands or salient item features differed between study and practice, context should change and, if the study context was not reinstated prior to the final test, RIF be enhanced. The context variations did not affect RIF, however, which disagrees with the context account.

In another recent study, Buchli, Storm, and Bjork (2016) examined RIF-like forgetting, adopting Delaney et al.'s (2010) near- vs. far-imagination technique as a means to manipulate amount of induced context change between study and restudy practice. The results of three recall experiments were reported, in each of which three restudy conditions were employed: a standard restudy condition; a restudy-plus-near-imagination condition, in which the restudy condition was preceded by an imagination task, in which participants were asked to imagine a "near" event (e.g., vacation in home country); and a restudy-plus-far-imagination condition, in which the restudy condition was preceded by an imagination task, in which participants were asked to imagine a "far" event (e.g., vacation abroad). Each experiment also included a standard retrieval practice condition. As expected, Buchli et al. found typical RIF. However, amount of RIF-like forgetting did not vary across imagination conditions, which disagrees with the context account. Moreover, surprisingly, no effect of imagination task was found at all, which directly contrasts with the results of Jonker et al. (2013), and the recall findings of the present Experiments 2 and 3.

As emphasized by Buchli et al. (2016), the experiments conducted by Jonker et al. and Buchli et al. were highly

[5] Recognition latencies varied slightly across experiments (Experiment 1: 1.896 ms; Experiment 2: 1.656 ms; Experiment 3: 1.613 ms), which was mainly due to the fact that restudied items were recognized faster than retrieval practiced items. However, in none of the three experiments were latencies faster than in Verde and Perfect's self-paced condition.

[6] The assumption that memory search (of the practiced items) during retrieval practice leads people away from the study context - although the items' category cues are provided as retrieval cues and people try to recall items from the study context - also raises a conceptual problem for the account. This is because, at the same time, the account assumes that memory search (of the control items) at test reinstates the study context when the items' category cues are provided as retrieval cues. Thus, the same type of memory search is supposed to induce context drift in the one case and context reinstatement in the other.

similar, and besides a few subtle methodological differences, like differences in study time or distractor task, differed in testing format only. Whereas in Jonker et al. item recall at test was blocked by category, in Buchli et al. item recall was blocked by item type, i.e., all unpracticed items were recalled before the list's practiced items. Because the latter procedure can increase the likelihood that, when recalling the unpracticed items, participants reinstate the study context rather than the practice context as the context account suggests (see above), the difference in results between the two studies may indeed be attributed to testing format.

However, testing format cannot explain the difference in results between Buchli et al. (2016) and the present study, in which item recall was also blocked by item type but context effects were found. Notably, besides a few subtle methodological differences, the study by Buchli et al. and the present one differed in duration of the context change task. While Buchli et al. - like Jonker et al. - asked participants in each single experiment to take part in one imagination task for one minute only to induce context change, here we asked participants in each single experiment to take part in two successive imagination (or semantic generation) tasks for two minutes each (see Method sections above). Such prolonged task may have created more robust effects of context change and may thus be (one of) the reason(s) why we did find RIF-like forgetting in recall, whereas Buchli et al. did not. All this is speculative, and future work is required to examine in more detail the effects of testing format and duration of context change task for the presence of RIF-like forgetting in recall.

Does context change contribute to RIF?

The present recognition results together with the recall results reported in the previous paragraphs challenge the context account of RIF as a full explanation of RIF. However, doing so, they do not rule out that context change may contribute to RIF. Indeed, to date at least three findings may be regarded as evidence that context change can contribute to RIF: (i) the finding that not only competitive but also noncompetitive retrieval practice can induce recall impairment for unpracticed items (Jonker & MacLeod, 2012; Raaijmakers & Jakab, 2012); (ii) the finding that also restudy when preceded by context change can induce recall impairment (Jonker et al., 2013; present Experiments 2 and 3); and (iii) the finding that reinstatement of the study context before test can reduce or even eliminate the recall impairment (Jonker et al., 2013). At least two of the three findings may not only be explained by context change.

Indeed, the presence of RIF with noncompetitive retrieval practice that Jonker et al. (2013) explained through context factors has also been explained by blocking processes. Raaijmakers and Jakab (2012) showed that retrieval of the category labels when the exemplars are provided intact as retrieval cues (e.g., _ - *ginger*) can impair recall of the unpracticed items and attributed the finding to strengthening of the cue-item associations during practice and blocking at test. Jonker and MacLeod (2012) employed subordinate generation (e.g., *dog* - _) as noncompetitive retrieval practice, in which the category exemplar was presented intact and participants were instructed to generate a type of *dog*, such as *beagle*. Such subordinate generation, however, impaired recall of the unpracticed items only when there was simultaneous retrieval of the category labels. Following Raaijmakers and Jakab, this raises the possibility that the RIF finding reported in Jonker and MacLeod was not mediated by context change but rather was caused by the strengthening of the cue-item associations during practice and blocking at test.

Blocking may also explain why restudy with preceding context change can cause RIF-like forgetting in recall (Jonker et al., 2013; present Experiments 2 and 3). This explanation relies on studies which showed that, after a change in internal context, item encoding can be improved. Corresponding evidence has been reported in context-dependent forgetting (Pastötter, Bäuml, & Hanslmayr, 2008) and multiple-list learning (Pastötter et al., 2011), although to date such improved encoding has been demonstrated for firstly studied items only. If context change also improved encoding of restudied items, then restudy after context change may enhance the strengthening of the associations of the items to their category label, induce blocking at test, and thus create RIF-like forgetting in recall.[7]

While two of the three findings mentioned above may thus be explained by both context change and blocking, Jonker et al.'s (2013) finding that both RIF and RIF-like forgetting can be eliminated when study context is reinstated at test cannot easily be reconciled with the blocking account, and it is also inconsistent with inhibition. Indeed, within the blocking account, there is no reason why reinstatement of the study context should eliminate the blocking effect of the practiced items; similarly, within the inhibition account, there is no reason why reinstatement of the study context should eliminate the inhibition of the unpracticed items. Likely, the finding provides the best current evidence for a possible contribution of context change to RIF. Future work should therefore try to replicate the result and examine its robustness and generalizability.

Multiple-factor accounts of RIF

Although most research on RIF during the past two decades was guided by the assumption that a single mechanism mediates RIF, there has also been research suggesting that two mechanisms may contribute to RIF, namely inhibition *and* blocking (e.g., Anderson & Levy, 2007; Aslan & Bäuml, 2010; Grundgeiger, 2014;

[7] The finding that restudy with preceding context change can cause RIF-like forgetting in recall may also be attributed to inhibition. Such an inhibition explanation would have to assume that (i) restudy triggers recognition processes that are more demanding in the presence than the absence of preceding context change, and (ii) such recognition can induce forgetting of unpracticed items, at least if recognition is demanding (e.g., Maxcey & Woodman, 2014). However, if inhibition mediated this form of RIF-like forgetting, the forgetting should not be restricted to recall but generalize to item recognition, which contrasts with the present results (see Experiments 2 and 3).

Rupprecht & Bäuml, 2016; Storm & Levy, 2012). Such a two-factor account posits that inhibition operates during retrieval practice and, in addition, blocking may arise during the final test. Importantly, inhibition is supposed to induce a retrieval-specific reduction in the unpracticed items' memory representation, observable over a wide range of memory tests. In contrast, blocking is proposed to play a role primarily in tests, in which item-specific cues are reduced, and to be largely eliminated in item recognition, in which the items themselves are presented as cues. Consequently, even though both inhibition and blocking may play a role in RIF in general, the particular test format should influence the relative contribution of the two mechanisms.

This two-factor account is consistent with a wide range of RIF findings, like, for instance, the presence of RIF in both recall and item recognition, the presence of RIF-like forgetting in recall but not in item recognition, the presence of RIF with noncompetitive retrieval practice, and the presence of RIF-like forgetting with restudy preceded by context change (for a more detailed discussion, see Rupprecht & Bäuml, 2016). All this holds while there are also challenges to this account, like Jonker et al. (2013) intriguing finding that both RIF and RIF-like forgetting may be eliminated when study context is reinstated at test. Because this latter finding points to an additional role of context factors in RIF, the question arises of whether not only inhibition and blocking but also context change may contribute to RIF, that is, whether the current 2-factor account of RIF may need revision towards a more general 3-factor account.

Future work, both empirical and theoretical, is required to address this important issue. Empirically, such work may examine whether effects of restudy with preceding context change arise in recognition tests that rely strongly, or even exclusively, on recollection judgments. For instance, in situations, in which distractors are highly similar to targets - like in plurality discrimination, where the lures are plurality reversed versions of the targets (e.g., Hintzman & Curran, 1994) - participants are forced to rely on recollection of specific discriminative details in order to respond correctly, and finding detrimental effects of restudy with context change in such situation would indicate that context change can contribute to RIF (for a related result in list-method directed forgetting, see Sahakyan, Waldum, Benjamin, & Bickett, 2009). Theoretically, future work may investigate in more detail exactly which mechanisms are necessary to explain the presence of RIF and RIF-like forgetting in different testing formats, and which set of mechanisms may be sufficient to explain the full range of (current) RIF findings. The two-factor account of RIF proposed above, which includes inhibition and blocking as core mechanisms, may serve as a useful baseline model for such theoretical analysis.

Conclusions

The context account of RIF claims that the effects of retrieval practice on unpracticed items can be simulated by restudy trials when these trials are preceded by context change. It therefore predicts that not only retrieval practice but also restudy preceded by context change should reduce both recall and recognition of unpracticed items. Testing this prediction, we found that retrieval practice impairs both recall and recognition of unpracticed items, whereas restudy preceded by context change impairs recall but leaves recognition of the items unaffected. Restudy with preceding context change thus cannot simulate RIF, which challenges the context account as a full explanation of RIF. Future work is warranted to examine in more detail the exact role of inhibition, blocking, and context change in RIF.

Authors' note

This research is part of J. Rupprecht's dissertation. We thank M.C. Anderson, T.R. Jonker, and L. Sahakyan for their suggestions on a previous version of the manuscript.

References

Anderson, M. C. (2003). Rethinking interference theory: Executive control and the mechanisms of forgetting. *Journal of Memory and Language, 49*, 415–445.

Anderson, M. C., Bjork, R. A., & Bjork, E. L. (1994). Remembering can cause forgetting: Retrieval dynamics in long-term memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 20*, 1063–1087.

Anderson, M. C., Bjork, E. L., & Bjork, R. A. (2000). Retrieval-induced forgetting: Evidence for a recall-specific mechanism. *Psychonomic Bulletin and Review, 7*, 522–530.

Anderson, M. C., & Levy, B. J. (2007). Theoretical issues in inhibition: Insights from research on human memory. In A. S. Benjamin (Ed.), *Successful remembering and successful forgetting: A festschrift in honor of Robert A. Bjork* (pp. 107–132). New York, NY: Psychology Press.

Anderson, M. C., & Spellman, B. A. (1995). On the status of inhibitory mechanisms in cognition: Memory retrieval as a model case. *Psychological Review, 102*, 68–100.

Aslan, A., & Bäuml, K.-H. T. (2010). Retrieval-induced forgetting in young children. *Psychonomic Bulletin and Review, 17*, 704–709.

Bäuml, K.-H. (2002). Semantic generation can cause episodic forgetting. *Psychological Science, 13*, 357–361.

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

Bäuml, K.-H., & Kuhbandner, C. (2007). Remembering can cause forgetting – But not in negative moods. *Psychological Science, 18*, 111–115.

Bäuml, K.-H., Pastötter, B., & Hanslmayr, S. (2010). Binding and inhibition in episodic memory – Cognitive, emotional, and neural processes. *Neuroscience and Biobehavioral Reviews, 34*, 1047–1054.

Bäuml, K.-H. T., & Samenieh, A. (2012). Selective memory retrieval can impair and improve retrieval of other memories. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 38*, 488–494.

Bodner, G. E., & Lindsay, D. S. (2003). Remembering and knowing in context. *Journal of Memory and Language, 48*, 563–580.

Bodner, G. E., & Richardson-Champion, D. D. L. (2007). Remembering is in the details: Effects of test-list context on memory for an event. *Memory, 15*, 718–729.

Buchli, D. R., Storm, B. C., & Bjork, R. A. (2016). Explaining retrieval-induced forgetting: A change in mental context between the study and restudy practice phases is not sufficient to cause forgetting. *The Quarterly Journal of Experimental Psychology, 69*, 1197–1209.

Ciranni, M. A., & Shimamura, A. P. (1999). Retrieval-induced forgetting in episodic memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 25*, 1403–1414.

Craik, F. I. M., & Schloerscheidt, A. M. (2011). Age-related differences in recognition memory: Effects of materials and context change. *Psychology and Aging, 26*, 671–677.

Delaney, P. F., Sahakyan, L., Kelley, C. M., & Zimmerman, C. A. (2010). Remembering to forget: The amnesic effect of daydreaming. *Psychological Science, 21*, 1036–1042.

Divis, K. M., & Benjamin, A. S. (2014). Retrieval speeds context fluctuation: Why semantic generation enhances later learning but hinders prior learning. *Memory and Cognition, 42*, 1049–1062.

Dunn, J. C. (2004). Remember-know: A matter of confidence. *Psychological Review, 111*, 524–542.

Godden, D. R., & Baddeley, A. D. (1975). Context-dependent memory in two natural environments: On land and underwater. *British Journal of Psychology, 66*, 325–331.

Godden, D. R., & Baddeley, A. D. (1980). When does context influence recognition memory? *British Journal of Psychology, 71*, 99–104.

Gómez-Ariza, C. J., Lechuga, M. T., Pelegrina, S., & Bajo, M. T. (2005). Retrieval-induced forgetting in recall and recognition of thematically related and unrelated sentences. *Memory and Cognition, 33*, 1431–1441.

Grundgeiger, T. (2014). Noncompetitive retrieval practice causes retrieval-induced forgetting in cued recall but not in recognition. *Memory and Cognition, 42*, 400–408.

Hicks, J. L., & Starns, J. J. (2004). Retrieval-induced forgetting occurs in tests of item recognition. *Psychonomic Bulletin and Review, 11*, 125–130.

Hintzman, D. L., & Curran, T. (1994). Retrieval dynamics of recognition and frequency judgments: Evidence for separate processes of familiarity and recall. *Journal of Memory and Language, 33*, 1–18.

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*, 112–127.

Jonker, T. R., & MacLeod, C. M. (2012). Retrieval-induced forgetting: Testing the competition assumption of inhibition theory. *Canadian Journal of Experimental Psychology, 66*, 204–211.

Jonker, T. R., Seli, P., & MacLeod, C. M. (2013). Putting retrieval-induced forgetting into context: An inhibition-free, context-based account. *Psychological Review, 120*, 852–872.

Kelley, R., & Wixted, J. T. (2001). On the nature of associative information in recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 27*, 701–722.

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* (pp. 171–189). New York, NY: Psychology Press.

Koessler, S., Engler, H., Riether, C., & Kissler, J. (2009). No retrieval-induced forgetting under stress. *Psychological Science, 20*, 1356–1363.

Macmillan, N. A., & Creelman, C. D. (2004). *Detection theory: A user's guide* (2nd ed.). London, NJ: Lawrence Erlbaum Assoc Inc..

Mannhaupt, H.-R. (1983). Produktionsnormen für verbale Reaktionen zu 40 geläufigen Kategorien. *Sprache und Kognition, 2*, 264–278.

Maxcey, A. M., & Woodman, G. F. (2014). Forgetting induced by recognition of visual images. *Visual Cognition, 22*, 789–808.

Moosbrugger, H., Oehlschlägel, J., & Steinwascher, M. (2011). *Frankfurter Aufmerksamkeits-Inventar 2*. Bern: Huber.

Murayama, K., Miyatsu, T., Buchli, D., & Storm, B. C. (2014). Forgetting as a consequence of retrieval: A meta-analytic review of retrieval-induced forgetting. *Psychological Bulletin, 140*, 1383–1400.

Parks, C. M., & Yonelinas, A. P. (2008). Theories of recognition memory. In H. L. Roediger, III (Ed.). *Cognitive psychology of memory of learning and memory: A comprehensive reference* (Vol. 2, pp. 389–416). Oxford: Elsevier.

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*, 977–982.

Pastötter, B., Bäuml, K.-H., & Hanslmayr, S. (2008). Oscillatory brain activity before and after an internal context change – Evidence for areset of encoding processes. *NeuroImage, 43*, 173–181.

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

Raaijmakers, J. G. W., & Jakab, E. (2012). Retrieval-induced forgetting without competition: Testing the retrieval specificity assumption of inhibition theory. *Memory and Cognition, 40*, 19–27.

Raaijmakers, J. G. W., & Jakab, E. (2013). Rethinking inhibition theory: On the problematic status of the inhibition theory for forgetting. *Journal of Memory and Language, 68*, 98–122.

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

Rupprecht, J., & Bäuml, K.-H. T. (2016). Retrieval-induced forgetting in item recognition: Retrieval specificity revisited. *Journal of Memory and Language, 86*, 97–118.

Sahakyan, L., & Hendricks, H. E. (2012). Context change and retrieval difficulty in the list-before-last paradigm. *Memory and Cognition, 40*, 844–860.

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

Sahakyan, L., Waldum, E. R., Benjamin, A. S., & Bickett, S. (2009). Where is the forgetting with list-method directed forgetting in recognition? *Memory and Cognition, 37*, 464–476.

Scheithe, K., & Bäuml, K.-H. (1995). Deutschsprachige Normen für Vertreter von 48 Kategorien. *Sprache und Kognition, 14*, 39–43.

Shiffrin, R. M. (1970). Memory search. In D. A. Norman (Ed.), *Models of human memory* (pp. 374–447). New York: Academic Press.

Shivde, G., & Anderson, M. C. (2001). The role of inhibition in meaning selection: Insights from retrieval-induced forgetting. In D. Gorfein (Ed.), *On the consequences of meaning selection: Perspectives on resolving lexical ambiguity* (pp. 175–190). Washington, DC: American Psychological Association.

Smith, S. M., Glenberg, A., & Bjork, R. A. (1978). Environmental context and human memory. *Memory and Cognition, 6*, 342–353.

Soares, J. S., Polack, C. W., & Miller, R. R. (2016). Retrieval-induced versus context-induced forgetting: Does retrieval-induced forgetting depend on context shifts? *Journal of Experimental Psychology: Learning, Memory, and Cognition, 42*, 366–378.

Spitzer, B. (2014). Finding retrieval-induced forgetting in recognition tests: A case for baseline memory strength. *Frontiers in Psychology, 5*, 1102.

Spitzer, B., & Bäuml, K.-H. (2007). Retrieval-induced forgetting in item recognition: Evidence for a reduction in general memory strength. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 33*, 863–875.

Storm, B. C., Bjork, E. L., & Bjork, R. A. (2007). When intended remembering leads to unintended forgetting. *The Quarterly Journal of Experimental Psychology, 60*, 909–915.

Storm, B. C., & Levy, B. J. (2012). A progress report on the inhibitory account of retrieval-induced forgetting. *Memory and Cognition, 40*, 827–843.

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*, 1392–1399.

Veling, H., & van Knippenberg, A. (2004). Remembering can cause inhibition: Retrieval-indiced inhibition as cue independent process. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 30*, 315–318.

Verde, M. F. (2013). Retrieval-induced forgetting in recall: Competitor interference revisited. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 39*, 1433–1448.

Verde, M. F., & Perfect, T. J. (2011). Retrieval-induced forgetting in recognition is absent under time pressure. *Psychonomic Bulletin and Review, 18*, 1166–1171.

Wimber, M., Alink, A., Charest, I., Kriegeskorte, N., & Anderson, M. C. (2015). Retrieval induces adaptive forgetting of competing memories via cortical pattern suppression. *Nature Neuroscience, 18*, 582–589.

Wixted, J. T. (2007a). Dual-process theory and signal-detection theory of recognition memory. *Psychological Review, 114*, 152–176.

Wixted, J. T. (2007b). Signal-detection theory and the neuroscience of recognition memory. In J. S. Nairne (Ed.), *The foundations of remembering: Essays in honor of H.L. Roediger, III* (pp. 67–82). New York: Psychology Press.

Wixted, J. T., & Stretch, V. (2004). In defense of the signal detection interpretation of remember/know judgments. *Psychonomic Bulletin and Review, 11*, 616–641.