Journal of Experimental Psychology:
Learning, Memory, and Cognition
2016, Vol. 42, No. 3, 366–378

0278-7393/16/$12.00 http://dx.doi.org/10.1037/xlm0000171

Retrieval-Induced Versus Context-Induced Forgetting: Does Retrieval-Induced Forgetting Depend on Context Shifts?

Julia S. Soares
University of California Santa Cruz

Cody W. Polack and Ralph R. Miller
State University of New York at Binghamton

Retrieval-induced forgetting (RIF) is the observation that retrieval of target information causes forgetting of related nontarget information. A number of accounts of this phenomenon have been proposed, including a context-shift-based account (Jonker, Seli, & Macleod, 2013). This account proposes that RIF occurs as a result of the context shift from study to retrieval practice, provided there is little context shift between retrieval practice and test phases. We tested both claims put forth by this context account. In Experiment 1, we degraded the context shift between study and retrieval practice by implementing a generative study condition that was highly similar to retrieval practice. We observed no degradation of RIF for these generated exemplars relative to a conventional study control. In Experiment 2, we conceptually replicated the finding of RIF following generative study, and tested whether context differences between each of the three phases affected the size of RIF. Our findings were again contrary to the predictions of the context account. Conjointly, the 2 experiments refute arguments about the potential inadequacy of our context shifts that could be used to explain either result alone. Overall, our results are most consistent with an inhibitory account of RIF (e.g., Anderson, 2003).

Keywords: retrieval-induced forgetting, context account, inhibition account, interference, reinstatement

Retrieval-induced forgetting (RIF) is a phenomenon that suggests that forgetting of some items is in part a consequence of remembering other items (Anderson, Bjork, & Bjork, 1994). Specifically, RIF is said to occur when retrieving some information causes forgetting of other information. A typical RIF experiment consists of four phases: study, retrieval practice, distractor task, and test. Most commonly, category-exemplar pairs (e.g., Color-Green, Color-Yellow, Bird-Crow, Bird-Robin) are presented during the study phase, and participants are asked to learn these pairs. During retrieval practice, they are commonly cued by category-exemplar stems (e.g., Color-Gr___) for half of the exemplars in half of the categories. The retrieval practice phase splits the categories into those that are retrieval practiced (Rp; e.g., Color) and those that not retrieval practiced (Nrp; e.g., Bird), and the Rp categories into exemplars (i.e., items) that were practiced (Rp+; Green) and exemplars that were not practiced (Rp–; Yellow). Following this, there is typically a distractor task or retention interval ranging from 2 to 20 min (although sometimes longer), and then a category-exemplar stem cued-recall test. Consistent with the testing effect (e.g., Bjork, 1975; Roediger & Karpicke, 2006; Storm, Friedman, Murayama, & Bjork, 2014), Rp+ items are more likely to be recalled than Nrp items, although practiced exemplars are typically compared with restudied items in testing effect research. RIF is manifest in the observation that Rp– items (i.e., those that were in a Rp category but never retrieval practiced) are less likely to be recalled than Nrp items (i.e., "Yellow" in the Color category being less likely recalled than "Crow" or "Robin" from the Bird category). More generally, RIF is observed when retrieval (or attempted retrieval) of target information causes forgetting of related nontarget information. Although this phenomenon is often studied with categories and exemplars, RIF is observed in a wide variety of tasks (see Storm et al., 2015, for a nonexhaustive review).

Arguably, the most widely supported account of RIF (see Storm & Levy, 2012) is an inhibitory account, which claims that competing items (Rp–; e.g., Color-Yellow) within a category are activated by the category cue (e.g., Color) during retrieval practice with Rp+ items (e.g., Green), and are then actively inhibited to reduce competition with the Rp+ item during the retrieval practice phase, resulting in item-specific inhibition that impairs later retrieval of the Rp– item at test (see Anderson, 2003). In this case, inhibition refers to "a suppression-type process directed at to-be-inhibited information for some adaptive purpose" (Bjork, 1989, p. 324). The adaptive purpose, in this case, is inhibition of distracting or interfering material (i.e., Rp– items during the retrieval practice phase) that competes for successful retrieval of relevant information (i.e., Rp+ items during the retrieval practice phase).

In a typical RIF paradigm, this inhibition presumably develops throughout retrieval practice (Kuhl, Dudukovic, Kahn, & Wagner, 2007; Wimber, Alink, Charest, Kriegeskorte, & Anderson, 2015)

This article was published Online First September 21, 2015.

Julia S. Soares, Department of Psychology, University of California Santa Cruz; Cody W. Polack and Ralph R. Miller, Department of Psychology, State University of New York at Binghamton.

This research was supported by National Institute of Mental Health Grant 33881. The authors thank Michael Anderson, Gonzalo Miguez, James Neely, Sarah O'Hara, and two anonymous reviewers for their comments on an earlier version of this article.

Correspondence concerning this article should be addressed to Ralph R. Miller, Department of Psychology, State University of New York–Binghamton, P.O. Box 6000, Binghamton, NY 13902-6000. E-mail: rmiller@binghamton.edu

and accumulates over multiple trials, because with each trial of practice, there is another opportunity for Rp– items to be inhibited (Storm, Bjork, & Bjork, 2008; Veling & van Knippenberg, 2004). In addition, this type of inhibition is posited to be cue independent, rather than a suppression or degradation of the dyadic association between category and exemplar. Evidence for this assertion comes from the observation that if different categories are used to cue recall of items at test than were studied (e.g., Fruit-Cherry is studied and Red-Ch___ appears at test), RIF is still observed in that Rp– items are less likely to be recalled than baseline Nrp items at test (e.g., Anderson & Spellman, 1995; cf. Williams & Zacks, 2001). This claim is also supported by studies that found RIF using a recognition test of items in which the category was not presented at test (Aslan & Bäuml, 2010, 2011; Grundgeiger, 2014; Hicks & Starns, 2004; Wimber et al., 2015) and by a compelling meta-analysis (Murayama, Miyatsu, Buchli, & Storm, 2014). These examples suggest that the specific item representation (e.g., Cherry) is suppressed and RIF does not result from interference or inhibition of the Fruit-Cherry association.

It has furthermore been proposed that this type of inhibition is retrieval-specific, in that additional study, as opposed to retrieval practice, of Rp+ items does not produce suppression of Rp– items (e.g., Anderson & Bell, 2001; Ciranni & Shimamura, 1999). This observation is also consistent with the inhibitory account's assertion that suppression of Rp– items does not depend on the associative strengthening of Rp+ items, as is Storm, Bjork, Bjork, and Nestojko's (2006) finding that RIF occurs even when it is made impossible for participants to successfully retrieve Rp+ items. Moreover, these observations challenge alternative, associative interference accounts of RIF, which posit that the strengthening of Rp+ items causes a deficit in Rp– items as a result of some sort of blocking or reallocation of mental resources during test (e.g., Camp, Pecher, & Schmidt, 2007; Jakab & Raaijmakers, 2009).

Although at least attempted retrieval is necessary for the occurrence of RIF, proponents of inhibitory accounts propose that retrieval alone is not sufficient. For RIF to occur, this retrieval must be also be competitive (i.e., activate Rp– exemplars during retrieval practice). Competition dependence is supported by studies that found no RIF when categories, rather than exemplars, are retrieval practiced (Anderson, Bjork, & Bjork, 2000, cf. Jonker & MacLeod, 2012; see also Wimber, Rutschmann, Greenlee, & Bäuml, 2009, for neuroimaging evidence of retrieval specificity). Moreover, RIF has not typically been observed when exemplars are only restudied rather than practiced (e.g., Anderson & Bell, 2001; Anderson, Bjork, & Bjork, 2000; Bäuml, 2002; Ciranni & Shimamura, 1999). Both restudy and category retrieval during the retrieval practice phase strengthen the category-cue association but presumably do not promote competition between exemplars within the category. In addition, Chan, Erdman, and Davis (2015) found attenuated RIF when Rp– items were presented for study only after retrieval practice (low competition) relative to when Rp– items were presented before retrieval practice (high competition), supporting the notion of competition dependence. As such, the inhibitory account predicts that nontarget items must compete for retrieval during retrieval practice for forgetting of these nontargets to be observed. To summarize, the inhibitory account posits that RIF is cue independent, retrieval specific, strength independent, and competition dependent. Additionally, this account assumes that inhibition is established during retrieval practice.

Although the inhibitory account of RIF has been well supported in the literature (e.g., Murayama et al., 2014; Storm & Levy, 2012; but see Raaijmakers & Jakab, 2013, and Verde, 2012), some studies have failed to support various assumptions of this account. For instance, the competition-dependence assumption (as well as retrieval specificity, which is presumably a precursor for competition), has faced some concerns. RIF does not seem to always require competition for retrieval (Jonker & MacLeod, 2012; Jonker et al., 2013; Ortega-Castro & Vadillo, 2013; Raaijmakers & Jakab, 2012), and the nature of competition according to the inhibitory account is rather vague, which makes it difficult for the inhibitory account to make clear predictions. Others have found that observations supporting cue independence are difficult to replicate (Jonker & MacLeod, 2012; Williams & Zacks, 2001). Still other researchers have observed attenuation of RIF when using item-specific independent cues argued to control most effectively for associative interference (Camp et al., 2007).

Jonker et al. (2013) proposed an alternative account, hereafter called the *context account*, to explain RIF in the absence of either an inhibitory mechanism or differences between Rp + and Rp– in strength of association to the category. The account proposes two tenets. Tenet 1 states that in order for RIF to be observed, a context shift must occur between the study phase and retrieval practice. Tenet 2 proposes that the final test must contextually resemble retrieval practice more so than the test resembles initial study. Specifically, the context account predicts "RIF will occur on the final test only when the practice context is reinstated for the practiced categories and when the study context is reinstated for the NRP categories," (Jonker et al., 2013, p. 855). The context shift created by switching from a study procedure (in which the full category and item are presented) to a retrieval practice procedure (in which cues are presented and the subject must retrieve the item) establishes multiple contexts for the Rp category. Jonker and colleagues suggest that presentation of a category at test constitutes selective reinstatement of the context in which the category was last presented along with the items experienced in that context. Thus, at test, Rp+ items benefit from activation of the retrieval practice phase (as they were actually presented during that phase), whereas Rp– items do not. Relative to Rp– items, Nrp items also benefit from reinstatement because these categories were previously presented only during study. Consequently, presentation of these categories reinstates the study phase, which accounts for superior recall of Nrp items relative to Rp– items.

As such, the context account does not predict cue independence, because the associated category must be presented in the final test to reinstate the retrieval context for practiced categories. Nor does the account posit that RIF is necessarily retrieval specific, because the processes assumed to cause the effect occur during the final test. Instead, the account claims that RIF-like effects are the result of the final test context reinstating the category's most recent and otherwise procedurally similar encounter, which may be produced by features other than retrieval processing. (Note that we say "RIF-like" because, according to this account, it is not retrieval of the Rp+ items per se, but presentation of the Rp category at test that produces the deficit in recalling Rp– items relative to Nrp items.) In fact, Experiments 1 and 2 reported by Jonker and colleagues (2013) did not employ a retrieval practice paradigm at all. Retrieval processing is considered merely one method of inducing a context shift between study and practice.

Context has often been shown to have a critical influence on memory performance. In a classic experiment, Godden and Baddeley (1975) asked divers to study lists of words either on land or under water, and then tested them in either the context in which they studied or the alternative context. Those subjects who were tested in the same context as they were trained recalled more items. A variety of changes with respect to physical context can influence memory, including background color (Dulsky, 1935), modality of list presentation, time of day, location (S. M. Smith, Glenberg, & Bjork, 1978), and appearance of the study room—although these effects are often smaller when stronger retrieval cues are given at test (e.g., S. M. Smith, 1979).

In addition to spatial and temporal contexts, the memory literature describes similar effects of internal context. For example, mood can provide an internal context that can modulate memory retrieval (e.g., Bower, 1981; E. Eich & Metcalfe, 1989). Drug states can also provide internal context cues (e.g., E. J. Eich, Weingartner, Stillman, & Gillin, 1975; Goodwin, Powell, Bremer, Hoine, & Stern, 1969; Overton, 1971). Moreover, the specific task participants are asked to perform during training and at test can influence success on a memory test. Subjects better remember items tested using a task that matches the task that was practiced (e.g., Morris, Bransford, & Franks, 1977; Tulving & Thomson, 1973). Particularly relevant to Jonker et al.'s (2013) context account is the finding that shifts from more passive processes, such as study and recognition to active retrieval (like the shift from study to retrieval practice in a typical RIF paradigm), can cause an internal context shift that influences memory performance (Jang & Huber, 2008; Sahakyan & Hendricks, 2012).

Because the context account is relatively new, it has yet to be tested extensively in the literature. However, Miguez, Mash, Polack, and Miller (2014) found no decrease in RIF as a result of a temporal and spatial context change between retrieval practice and test when the context of test matched that of study. This finding challenges the context account, which would predict no RIF when the context of study is active during test. If the context of study is active during test, this would not allow for selective contextual reinstatement of the practiced categories, because all items are presented for study, so the account predicts no differences in recall of item types in practiced and nonpracticed categories. However, the salience of these environmental context shifts was not independently assessed, so proponents of the context account might argue that the changes in context were not sufficient to constitute a substantive context shift for participants. In contrast, components of the cognitive or internal context—and particularly changes in presentation style to which participants are exposed in the different phases in the standard RIF paradigm—are apt to be more salient than spatial and temporal features because they are more critical components of the task. So the context of retrieval practice (rather than study) may have still been active at test, even if the environmental context features at test resembled those of study, because test itself resembled retrieval practice more than study.

The experiments accompanying Jonker and colleagues' (2013) context account provide some compelling evidence for this new account of RIF, but stray from the typical experimental RIF paradigm. Their Experiments 1 and 2 deviate most notably in that they did not use any of the retrieval practice procedures that are ordinarily employed in demonstrations of RIF. Instead, these experiments employed additional study for some items rather than retrieval practice. Presumably, this was intended to reduce retrieval processing as a similar internal context between practice and testing. Jonker et al. also manipulated context shifts using an imagination task and videos. Although these tasks were used effectively to create context shifts between phases, they deviate from the mechanisms that the context account argues control context shifts in a standard RIF paradigm—the differences in task demands and presentation style between phases.

In the present series, we attempted to assess predictions based on the context account concerning the two tenets of the account as well as the underlying mechanism that Jonker et al. (2013) propose by manipulating the context of each phase with respect to processing and presentation style. In Experiment 1, we tested Tenet 1 by employing two different Phase 1 initial learning procedures that were similar or different from the procedure used during retrieval practice. This was achieved by using generative study for some items, which was intended to match Phase 1 to the dominant procedural similarity shared by retrieval practice and test, that is, retrieval processing. The test procedure was always similar to the retrieval practice procedure, thereby fulfilling Tenet 2 of the context account, in that there was never greater similarity between test and initial study than between test and retrieval practice. Temporal proximity would have always favored similarity between test and retrieval practice. Hence, based on Tenet 1, the context account anticipates that RIF would be greater when the procedures used during initial learning in Phase 1 (hereafter called the *learning phase*) and retrieval practice in Phase 2 (henceforth called the *retrieval practice phase*) were different relative to when they were more similar. According to the context account, minimizing the differentiation between learning and retrieval practice should reduce the Rp category cues' selective ability to reinstate the context of retrieval practice by making these two phases difficult to differentiate. The inhibitory and the context-based accounts make divergent predictions here, in that according to an inhibitory account it should make little difference whether the procedures of initial learning and retrieval practice are the same or different.

In Experiment 2, we tested specific predictions made by the context account based on varying the context match or mismatch between the learning, retrieval practice, and test phases. We not only assessed Tenet 2 of the account (the role of learning/test-context match), but again tested Tenet 1 (the role of learning/retrieval-practice-context match) and mechanistic predictions of the context account (i.e., the role of learning/test-context match).

Experiment 1

We first addressed Tenet 1 of the context account—that the contexts of study and retrieval practice must differ. This context shift is necessary to ensure reinstatement of the retrieval practice context at test is differentially providing a benefit to items that were Rp in Phase 2 relative to nonpracticed items in the Rp category that were only studied during Phase 1. If the study and retrieval practice contexts are similar, it would be difficult to differentially reinstate one as opposed to the other at test, effectively because the subject would have difficulty telling the two phases apart. Therefore, reducing the contextual distinction between study and retrieval practice should diminish RIF. Typically in studies of RIF, this context shift occurs as a result of the nature of the different tasks in Phases 1 and 2. Learning (i.e., "study") is

usually a relatively passive task whereby participants are asked to read and remember a series of word pairs presented to them. Retrieval practice is necessarily a more active process of searching for a particular item in memory that had previously been paired with the category presented during retrieval practice and fits the stem letters. In Experiment 1, the critical condition presented participants with a forced generate task in place of the conventional study task during Phase 1. This was done with the intent of attenuating the context shift (particularly in terms of retrieval processing) between learning in Phase 1 and retrieval practice in Phase 2. According to the context account, this change to a generate task in Phase 1 should result in (at least) reduced RIF compared with a typical RIF group that experiences a conventional study task in Phase 1.

Method

Participants. A total of 131 participants were recruited from the State University of New York at Binghamton's (hereafter, SUNY-Binghamton) undergraduate subject pool for this experiment. The data from 11 participants were removed from the analysis based on our elimination criteria (described in Results and Discussion), leaving 120 participants (69 female), ages 18 to 29 years, for our final analysis.

Design. The study used a 2 (generate vs. study during Phase 1) × 2 (retrieval practice vs. no retrieval practice during Phase 2) design, fully within subjects (see Table 1). The experiment consisted of four phases (learning [i.e., conventional study or generate], retrieval practice, distractor task, and test). This sequence of four phases was administered twice consecutively, each time with entirely different Rp and Nrp categories. Thus, each participant experienced each of the two learning conditions (i.e., study and generate), counterbalanced across subjects for order. The use of two Rp and Nrp categories within each learning condition was intended to double the amount of resultant data, which was then averaged for each type of category within each condition. "Generate" during Phase 1 Learning refers to presentation of category-exemplar stem pairs (e.g., Vegetable-Broc__) with instructions to complete the stem presented. The first syllable (according to Dictionary.com) of each item in this condition was provided to direct generation of a particular exemplar, unless the first syllable had fewer than three letters, in which case the first three letters were provided instead. Participants had to generate and type the complete exemplar. "Study" during learning refers to presentation of the category and complete exemplar for participants to read.

Practiced (Rp) categories had half ($n = 4$) of their category-exemplar stems (i.e., first syllables) Rp during Phase 2. None of the items in nonpracticed (Nrp) categories were Rp during Phase 2. Only the four lower frequency items in any practiced category were Rp (Rp+); that is, the four higher frequency items were not Rp (Rp–). This was done with the intention of making our paradigm as sensitive as possible to RIF, as prior research has suggested that using higher frequency items as Rp– targets leads to more RIF presumably because they are more competitive during retrieval practice (see Anderson et al., 1994). Moreover, to avoid item frequency confounding retrieval practice, Nrp items were also split into those with high taxonomic strength (Nrp[–]) for comparison with Rp– items and low taxonomic strength (Nrp[+]) for comparison with Rp+ items.

Each of the two initial learning conditions (i.e., study and generate) was presented as an individual RIF experiment (i.e., a sequence of four phases) with two practiced and two nonpracticed categories. Each "miniexperiment" used entirely different categories; thus, there were a total of eight experimental categories (four per learning condition), each containing eight exemplars. The Study sequence and the Generate sequence (each consisting of four phases), were presented consecutively, with order of study and generate counterbalanced across participants. Category-exemplar pair order was randomized within trial blocks during all conditions. During learning and test, each block contained one category-exemplar pair from each category, so there were four sets in each block, and a total of eight blocks. During retrieval practice, because only half of the items in half of the categories were presented, there were four blocks with two items in each block. One filler category-item pair was placed between adjacent blocks to prevent items within the same category being presented sequentially. Two filler category-item pairs also preceded each phase to acquaint participants with the different tasks as well as attenuate primacy effects. Two filler category-item pairs at the end of learning and retrieval practice were presented to attenuate recency effects.

Order for generate versus study was counterbalanced, and this was comanipulated with category assignment to condition (four

Table 1
Experiment 1 Design (Representative Categories)

Condition	Phase 1 learning (32 category-exemplar sets)	Phase 2 retrieval practice (Eight category-exemplar stem sets)	Phase 4 test (32 category-exemplar stem sets)	Context account (anticipated outcome)
Study	Music-Classical	Music-Gos___	Music-C___	RIF
Rp	Music-Gospel		Music-G___	
Nrp	Vegetable-Potato		Vegetable-P___	
	Vegetable-Cabbage		Vegetable-C___	
Generate	Tree-Map___	Tree-Dog___	Tree-M___	No/LessRIF
Rp	Tree-Dog___		Tree-D___	
Nrp	Drug-Her___		Drug-H___	
	Drug-Nic___		Drug-N___	

Note. Simplified design for Experiment 1. Study items were presented intact during learning, whereas generate items had only the first syllable or first three letters presented. Rp = categories presented during retrieval practice; Nrp = categories not presented during retrieval practice; RIF = retrieval-induced forgetting.

counterbalancing conditions constituting an incomplete Latin square). That is, pairs of categories were yoked together and then counterbalanced across participants. Thus, there were a total of eight counterbalancing conditions.

Materials. Exemplars were chosen from lists found in Van Overschelde, Rawson, and Dunlosky (2004). Eight categories, each containing eight items, were selected for the experimental conditions (64 exemplars total). The exemplars ranged from five to 10 letters long, and two to four syllables. Each category was sorted by category typicality based on taxonomic frequencies reported by Van Overschelde and colleagues. The four filler categories were similarly selected. The distractor task was administered with pen and paper, and the rest of the experiment was administered using computers.

Procedure. The experiment consisted of four phases, administered twice (once for study and once for generate in Phase 1). The phases included a learning phase (study or generate), a retrieval practice phase, a distractor task phase, and a test phase. During the learning phase, in the study condition, intact category-exemplar pairs were presented for 7 s, with a 0.25-s interstimulus interval. In the generate condition, participants saw a category plus an incomplete stem consisting of the first syllable of the target item (or the first three letters, whichever was longer) and four underscores, and instructions to complete the incomplete word. In the generate, retrieval practice, and test phases, a gray box appeared in the middle of the screen below the category-exemplar pair stem. Whenever participants were expected to make a response, this box appeared, along with a note at the bottom of the screen that read "Please type the COMPLETE second word."

The initial instructions were identical for the study and generate conditions. Participants were informed that they would be participating in a computer-based memory experiment, and that they would be tested on the words seen during the experiment. They saw the following instructions:

> You will be asked to do a series of tasks that may require you to either READ, or GENERATE words. If you are presented with two full words (e.g., Spice—Cinnamon) you should read and concentrate on the words for the full time they are presented. If you are presented with a stimulus with an incomplete word (e.g., Spice—Cinn___), you should GENERATE the complete word (e.g., type "cinnamon"). A response box will appear whenever you are expected to make responses.

Retrieval practice followed, with the instruction "Please type the COMPLETE second word" on the bottom of each screen. Other than this, there were no additional instructions, so that in the generate condition, learning transitioned into retrieval practice without any change in presentation style or new instructions. In contrast, in the study condition, learning transitioned into retrieval practice with a clear change in presentation style and task demand (but still no break for instructions). All participants were presented with category-exemplar stems and asked to type the complete second word in the box that appeared on screen. Each Rp+ item was practiced three times (i.e., three blocks of Rp+ trials). Retrieval practice was identical to the generate condition of Phase 1.

The distractor task was a pencil-and-paper maze task stored in a folder at each desk. This task was administered for 1 min during each distractor phase. Participants returned to the same maze during each distractor phase. Participants were instructed to inform the experimenter if they finished the distractor task before the time allotted had elapsed; no participant did so.

Participants saw the following instructions at test: "Now you will be presented with a list of word pairs. In all cases the second word will be incomplete. You will have 10 seconds type the complete second word which has previously been shown." Test consisted of presentation of category-exemplar stems. These stems differed from the generative stems presented during generate and retrieval practice in that they contained only the first letter of the target exemplar (e.g., Vegetable-B____). This change was made to avoid ceiling effects for recall as well as allow for the scoring protocol described in Results and Discussion. As such, all items within a category had a unique first letter. High-frequency items (Rp– and Nrp[–]) were presented before any low-frequency items (Rp+ and Nrp[+]) to attenuate retrieval competition of the critical high-frequency items (Rp–) by prior testing on low-frequency items from the same category. They were permitted to backspace to fix errors in their response before submitting the response by pressing "Enter." Participants had 10 s to retrieve and respond to the presented stimulus, and following entry of their response and a 0.25-s intertrial interval, they were presented with the next category-exemplar stem.

Following their first test phase, participants were informed that they would not again be tested on the preceding category-exemplar pairs. Then the four phase sequence started over again with Phase 1 being switched between study and generate relative to the first four phase sequence. After their second test phase, participants were debriefed and thanked for their participation.

Results and Discussion

Participants who failed to make any correct responses during either test were eliminated. Because of this, data for 11 participants were replaced with 11 new participants, thereby maintaining counterbalancing of order (i.e., study vs. generate in Phase 1) and category assignment to condition. Responses were considered to be correct if they contained the stem used to initially cue generation (i.e., the first syllable of the target item). This protocol was used so that even if subjects generated an item that did not match the target, but did fit the generative stem, their response would still be counted as correct. We used this scoring criterion in an attempt to make study items more comparable to generate items and, at the same time, control for the selection bias that would have existed had we identified a specific word as a correct response. In the generate condition, it seemed possible that subjects would produce unexpected items that fit the stem, so we identified correct answers using the generative stem rather than fixating on how the participants "should" have responded.

Low-frequency items. The data were transformed into difference scores (Rp+ – Nrp[+]) reflecting the magnitude of a retrieval benefit for the low-frequency items (see Table 2 for means by item type). The difference in magnitude of the retrieval benefit between conditions indicated that this benefit was smaller for generated than studied items, $t(119) = 3.12$, $p = .002$, Cohen's $d = 0.57$, 95% confidence interval [CI] for d [0.20, 0.93] (see Figure 1). Context accounts of the testing effect suggest that there should be more of a retrieval benefit when there is a context shift between initial learning and retrieval practice (i.e., the study condition) compared with when this context shift is attenuated (i.e., the generate condition). With decreased ability to distinguish between

Table 2
Experiment 1 Descriptive Statistics

	Study		Generate	
	M	*SEM*	*M*	*SEM*
Rp+	.75	.02	.70	.02
Nrp[+]	.42	.02	.47	.02
Rp–	.48	.02	.56	.02
Nrp[–]	.52	.02	.62	.02

Note. Descriptive statistics for each item type in Experiment 1. "Study" indicates items that were studied during learning. "Generate" indicates items that were generated during learning. Rp+ = items that were practiced during Phase 2; Rp– = nontarget items in retrieval practiced categories; Nrp[+] = items belonging to nonpracticed categories with low taxonomic frequency; Nrp[–] = similar items with high taxonomic frequency; *SEM* = standard error of the mean.

retrieval practice and initial learning in the generate condition, the category presented at test is less able to preferentially reactivate the retrieval practice context relative to the learning context. That is, in the generate condition, initial learning and retrieval practice are more similar, so reinstatement of the retrieval practice context facilitates recall of items from the initial learning phase as well. Karpicke, Lehman, and Aue (2014) argue that a context shift between initial learning and retrieval practice allows subjects to restrict their search at test to items that were presented in the retrieval practice context, thereby reducing competition. The resulting decreased competition between practiced and nonpracticed items in the study condition may explain our observation of a larger testing effect in the study condition. Critically, the enhanced retrieval benefit that we observed in the study condition provides evidence that there was indeed a more salient context shift between study and retrieval practice than between generate and retrieval practice. Nevertheless, we did observe some retrieval practice benefit in both the study, $t(119) = 13.60$, $p < .001$, Cohen's $d = 2.48$, 95% CI [2.01, 2.96], and generate conditions, $t(119) = 10.58$, $p < .001$, Cohen's $d = 1.93$, 95% CI [1.50, 2.37].

High-frequency items. Participant data was similarly transformed to difference scores for the high-frequency items (Nrp[–] – Rp–) to assess the magnitude of RIF for each participant (see Table 2 for means by item type). We determined that overall RIF scores differed significantly from zero, $t(119) = 3.13$, $p = .002$, overall mean RIF effect = 0.05, standard error of the mean (*SEM*) = 0.02, Cohen's $d = 0.57$, 95% CI [0.21, 0.94], indicating that our procedures were able to produce a RIF effect overall. Contrary to the predictions of the context account, we observed no evidence of a significant difference between study RIF scores and generate RIF scores, $t(119) = 0.67$, $p = .503$, difference in *M*s = −0.02, 95% CI [−0.08, 0.04] (predictions of the context account are indicated in the positive direction; see Figure 2). Thus, were this difference significant, it would be in the direction opposite that predicted by the context account. As such, we computed a 95% CI in the direction predicted by the context account (i.e., that the mean RIF score for studied items is greater than the mean RIF score for generated items). With 95% confidence, the largest standardized mean difference (Cohen's *d*) that could be accounted for by the context account (i.e., on the positive end of the CI) was 0.10. That is, the context account describes, at most, a small effect—contrary to the claim that context alone is responsible for the RIF effect. Further support for this null finding was provided by a Bayesian *t* test analysis with odds of 8.79 in favor of the null hypothesis (Rouder, Speckman, Sun, Morey, & Iverson, 2009).

Although there was no detected difference in the magnitude of RIF between conditions, we wanted to be certain that RIF was reliable in both, so *t* tests comparing the size of each RIF effect with zero were conducted. We observed marginal RIF (indicated by lower recall of Rp– items than high-frequency Nrp items) in the study condition, $t(119) = 1.80$, $p = .073$, Cohen's $d = 0.33$, 95% CI [−0.03, 0.68], and statistically significant RIF in the generate

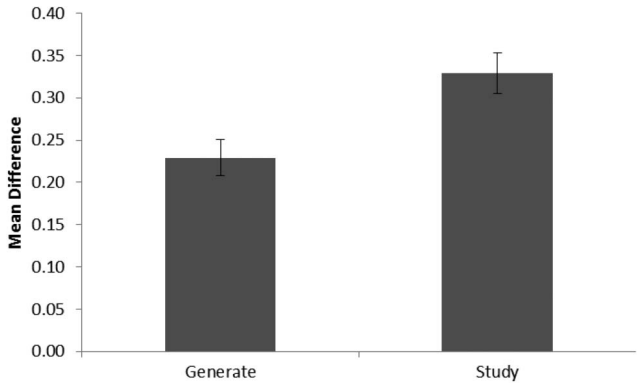

Figure 1. Mean difference scores reflecting the retrieval benefit of retrieval practice in low-frequency (Rp+ and Nrp[+]) items observed in Experiment 1. These difference scores reflect the proportion of retrieval practiced (Rp+) items relative to baseline items in comparable nonretrieval practiced categories (Nrp[+]; i.e., Rp+ – Nrp[+]). As such, positive values indicate more of a retrieval benefit. The conventionally studied condition is indicated by "Study," whereas the generative study condition is indicated by "Generate." Error bars indicate standard error of the means, but they should viewed cautiously, as the experiment was fully within subjects.

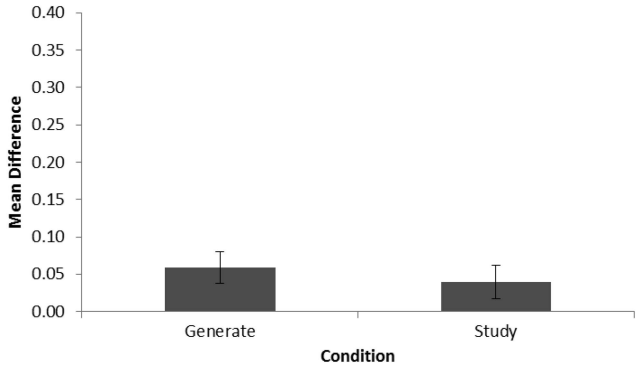

Figure 2. Mean difference scores reflecting the retrieval-induced forgetting in high-frequency (Rp– and Nrp[–]) items observed by condition in Experiment 1. These difference scores reflect the proportion of nontarget items in retrieval practiced categories (Rp–) relative to baseline items in comparable nonretrieval practiced categories (Nrp[–]; i.e., Nrp[–] – Rp–). As such, positive values indicate more retrieval-induced forgetting. The conventionally studied condition is indicated by "Study," whereas the generative study condition is indicated by "Generate." Error bars indicate standard error of the means, but they should viewed cautiously, as the experiment was fully within subjects.

condition, $t(119) = 2.80$, $p = .006$, Cohen's $d = 0.51$, 95% CI [0.14, 0.87].

Correlations. Notably, we observed a negative correlation between the magnitude of RIF and the magnitude of retrieval benefit (i.e., the testing effect) across conditions, $r(118) = -0.20$, $p = .002$, 95% CI [−0.32, −0.07]. This finding directly contradicts strength-based accounts of RIF, which posit that RIF depends on strengthening of Rp+ items (i.e., predicts a positive correlation). Here, we saw the opposite tendency (less RIF when individuals exhibited more of a retrieval benefit).

Overall, these results indicate that RIF does not depend on a context shift between learning and retrieval practice, as predicted by the context account. The context account predicts that in the generate condition, in which the context shift between Phases 1 and 2 is degraded, we should have observed at least an attenuated RIF effect compared with the study control. On the contrary, the interaction we observed, although not significant, was nominally in the opposite direction.

This experiment does not conceptually replicate the lack of RIF when the context shift between Phase 1 and Phase 2 is attenuated (i.e., Experiment 1 of Jonker et al., 2013). It has been well-demonstrated that RIF does not typically occur when items are restudied during Phase 2 rather than Rp (e.g., Anderson & Bell, 2001; Anderson et al., 2000; Bäuml, 2002; Ciranni & Shimamura, 1999). The results of the present Experiment 1 demonstrate that a shift from passive study to active retrieval practice is not necessary to demonstrate RIF. This result qualifies the context account by indicating that very subtle changes in context (i.e., the context shift between generation and retrieval practice) would have to qualify as context shifts for Tenet 1 to be satisfied in the present experiment.

Experiment 2

To further test the context account, Experiment 2 sought to conceptually replicate the finding of RIF using only the generative learning procedure of Experiment 1. This experiment was designed to address the possibility that the subtle context shift between generation and retrieval practice (effectively *regeneration*, as subjects were given no intervening instructions between these phases) was noticeable enough to constitute a salient context shift that may have allowed participants in Experiment 1 to differentiate between these phases and consequently differentially activate these two distinct contexts at test. To do so, we made generative study constant across conditions and attempted to manipulate context shifts separately by varying the display's background color, font color, font type, and the sequential versus simultaneous presentation of the generation cues. This experiment investigated the impact of learning/retrieval practice context match or mismatch (Tenet 1), retrieval practice/test context match or mismatch (Tenet 2), and learning/test-context match (critical to the context account's proposed underlying mechanism for RIF-like effects at test) on RIF.

Method

Participants. A total of 149 participants were recruited from SUNY-Binghamton's undergraduate subject pool for this experiment. Five subjects' data were not included in the analysis because they failed to meet the aforementioned criteria for inclusion. These subjects were replaced to maintain counterbalancing, so a total of 144 participants' data were included for analysis (109 female; ages 17 to 24 years).

Design. Procedures for each RIF cycle resembled those of Experiment 1's generate condition except as noted otherwise. The experiment consisted of four cycles, each of which consisted of a generative learning phase, retrieval practice phase, distractor task phase, and test phase. We used a fully within-subject factorial 2 (Context A vs. Context B during learning) × 2 (retrieval practice vs. no retrieval practice categories) × 2 (Context A vs. Context B during test) design (see Table 3). Context A consisted of presenting a category-exemplar stem pair in black text on a white background, and allowing Participants 10 s to generate (Phase 1) or retrieve (with generate possibly contributing; Phases 2 and 4) and type the full exemplar. Context B consisted of presenting the category alone (e.g., Fruit) for 1 s, then adding the first letter of the exemplar (e.g., Fruit-P____) for 1 s, and then showing the full

Table 3
Experiment 2 Design (Representative Categories)

Condition	Generative learning	Retrieval practice	Test	Context account prediction
AAA	Fruit-Ban___	Fruit-Ban___	Fruit-B___	Little/No RIF Tenet 1 violated
AAB	Fruit-Ban___	Fruit-Ban___	Fruit-___ Fruit-B___	No RIF Tenets 1 & 2 violated
BAA	Fruit-___ Fruit-B___ Fruit-Ban___	Fruit-Ban___	Fruit-B___	RIF
BAB	Fruit-___ Fruit-B___ Fruit-Ban___	Fruit-Ban___	Fruit-___ Fruit-B___	Little/No RIF Tenet 2 violated

Note. Simplified design of Experiment 2. In the Condition column, A and B represent two distinct presentation formats with the first of the three letters indicating the generative learning context, the second letter indicating the retrieval practice context, and the third letter indicating the test context. Context A refers to simultaneous (i.e., conventional) presentation of items in Times New Roman font, with black text on a white background. Context B refers to sequential presentation of items in Tahoma font, with white text on a black background. These contexts were actually counterbalanced across participants. The multiple entries in a cell (e.g., Fruit, Fruit-B___) indicate sequential presentation of first the category alone and then letters from the item.

stem, which was the first syllable according to Dictionary.com or the first three letters, whichever was longer (e.g., Fruit-Pap___). Participants were instructed to generate (Phases 1 and 2) or retrieve (Phases 2 and 4) and type the full exemplar (i.e., "Papaya") within 8 s of the response box being presented, which occurred only when the full stem was presented. Participants experienced each Learning × Test condition separately (a total of four conditions), counterbalanced for order. As such, each participant experienced four cycles of learning, retrieval practice, distractor task, and test. Each Learning × Test condition was assessed using four categories (two Rp and two Nrp), necessitating a total of 16 experimental categories used across the four condition cycles.

As previously stated, each of the four Learning × Test conditions was run in its entirety separately from the other three conditions. As such, four categories belonged to each condition (two Rp and two Nrp), with eight exemplars per category. Category assignment was comanipulated with Learning × Test condition order of presentation in the counterbalancing (i.e., an incomplete Latin square was used), whereas retrieval practice status of each category (Rp vs. Nrp) and context assignment (i.e., which conditions acted as Context A or Context B) was fully counterbalanced, necessitating 16 counterbalancing conditions.

Materials. Categories of exemplars were chosen from lists found in Van Overschelde et al. (2004). Sixteen categories, each containing eight items were selected (a total of 128 items). The exemplars ranged from five to 12 letters long, and contained one to four syllables. Additionally, eight filler categories were selected, four with eight items and four with four items (48 total).

Procedure. As indicated in Table 3, for each participant, this experiment consisted of four cycles, one for each Learning × Test condition, with each cycle consisting of a generative learning phase, a retrieval practice phase, a distractor task phase, and a test phase. For the Context A learning phase, category-exemplar stem sets (see Table 3) were presented intact for 7 s in black Times New Roman font text on a white screen. During the Context B learning phase, the category was first presented alone for 1 s, then the category-first-letter stem was presented for 1 s, and then, finally, the full stem was presented for 5 s. For Context B, the cues were presented in white Tahoma font text on a black background. In actuality, Contexts A and B were counterbalanced. Participants were not able to type until the full category-exemplar stem was presented.

In all phases, participants were instructed to press "Enter" in each phase to submit their responses. Following submission, there was a 0.25-s intertrial interval, and then participants were presented with the next stimulus. If they failed to respond, the next stimulus would be presented after the times specified.

During the retrieval practice phase, participants were presented with Rp+ category-exemplar stems in the same style as in the Context A generative learning phase. The distractor phase was the same as in Experiment 1. Again, no participant finished this task before the scheduled end of the distractor task.

Other than the shortened stems (first letter only), test proceeded as previously described for Contexts A and B, except that total presentation time before cutoff was elongated to 10 s per item. Because of this, the presentation style of Context B test was slightly altered in that category was presented alone for 1 s, and then the single letter stem was presented for 9 s. The categories and item stems presented in Context A were presented simultaneously for a total of 10 s.

Following the first, second, and third test phases, participants were informed that they would not again be tested on the words they had previously encountered, and that the next part of the experiment would involve entirely new word lists. After their fourth test phase, participants were debriefed and thanked for their participation.

Results

Responses were scored using the same protocol as Experiment 1.

Low-frequency items. Scores were transformed using the same protocol as was used for Experiment 1, resulting in a score reflecting the retrieval benefit demonstrated by each participant (Rp+ – Nrp[+]; see Table 3 for means by item type, and Figure 3 for mean retrieval benefit scores). First, we compared the pooled retrieval benefit scores across conditions with zero and found a robust retrieval benefit, $t(143) = 11.27$, $p < .001$, overall mean retrieval benefit = 0.11, $SEM = 0.02$, Cohen's $d = 2.06$, 95% CI [1.61, 2.50]. A 2 (learning/retrieval-practice-context match vs. mismatch) × 2 (retrieval-practice/test-context match vs. mismatch) within-subjects ANOVA was conducted. We detected no significant interaction between the factors, which is equivalent to the effect of learning/test-context match or mismatch on the size of a retrieval benefit, $F(1, 143) = 0.60$, $p = .442$. We also observed no reliable main effect of learning/retrieval-practice-context match, $F(1, 143) = 0.22$, $p = .638$, nor any significant effect of retrieval-practice/test-context match, $F(1, 143) = 0.002$, $p = .963$.

High-frequency items. Scores were transformed using the same protocol as was used for Experiment 1, resulting in a score reflecting RIF demonstrated by each participant (Nrp[–] – Rp–; see Table 4 for means by item type, and Figure 4 for mean RIF scores).

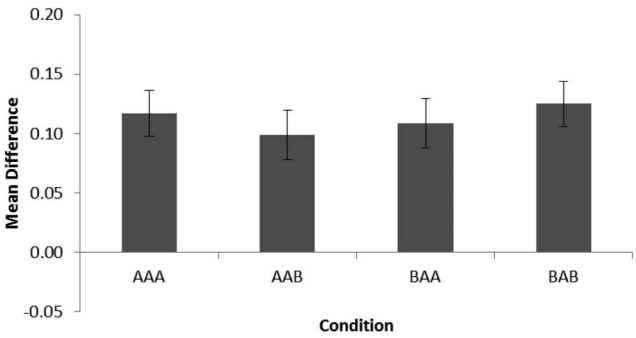

Figure 3. Mean difference scores reflecting the retrieval benefit in low-frequency (Rp + and Nrp[+]) items observed by condition in Experiment 2. These difference scores reflect the proportion of retrieval practiced (Rp+) items relative to baseline items in comparable nonretrieval practiced categories (Nrp[+]; i.e., Rp+ – Nrp[+]). As such, positive values indicate more of a retrieval benefit. The first letter indicates the context of learning, whereas the second letter signifies the context of retrieval practice, and the last letter indicates the context of test. These results indicate a robust effect of retrieval practice on retrieval practiced items (i.e., the testing effect). Error bars indicate standard error of the means, but they should viewed cautiously, as the experiment was fully within subjects. Note that the vertical scale used here differs from that of Figure 1.

Table 4
Experiment 2 Descriptive Statistics by Item Type

	AAA		AAB		BAA		BAB	
	M	*SEM*	*M*	*SEM*	*M*	*SEM*	*M*	*SEM*
Rp+	.59	.02	.57	.02	.55	.02	.60	.02
Nrp[+]	.47	.02	.47	.02	.45	.02	.47	.02
Rp–	.64	.02	.61	.02	.61	.02	.58	.02
Nrp[–]	.68	.02	.65	.02	.62	.02	.64	.02

Note. Descriptive statistics for each item type in Experiment 2. Conditions are indicated by the context presented in *learning* as the first letter, the context of *retrieval practice* as the second letter, and the context of *test* as the final letter. Rp+ = refers to items that were practiced during Phase 2; Rp– = items were nontarget items in retrieval practiced categories; Nrp[+] = items belonging to nonpracticed categories with low taxonomic frequency; Nrp[–] = similar items with high taxonomic frequency; *SEM* = standard error of the mean.

RIF scores (Nrp[–] – Rp–) were pooled across conditions and compared with zero; we observed reliable overall RIF, $t(143) = 3.78$, $p < .001$, overall mean RIF effect = 0.04, *SEM* = 0.02, Cohen's $d = 0.69$, 95% CI [0.29, 0.96]. A 2 (learning/retrieval-practice-context match vs. mismatch) × 2 (retrieval-practice/test-context match vs. mismatch) within-subjects ANOVA was conducted on the RIF scores. Results were also assessed for effects of order, with no main effect of order detected in the size of RIF, $F(1, 143) = 0.05$, $p = .984$, nor any interaction between order and condition, $F(1, 143) = 0.49$, $p = .883$.

The context account posits that RIF depends on the context match between learning and test, which allows for selective contextual reinstatement of Rp– and Nrp items. Contrary to this prediction, we did not observe a reliable interaction between learning/retrieval-practice context (match vs. mismatch) and retrieval-practice/test context (match vs. mismatch), which is equivalent to the effect of learning/test-context match of RIF, $F(1, 143) = 1.78$, $p = .185$, difference in *M*s = −0.05, 95% CI [−0.06, 0.01]. With 95% confidence in the direction predicted by the context account (e.g., more RIF in Conditions AAB and BAA relative to AAA and BAB), the largest standardized mean difference (*d*) predicted accurately by this account was 0.04, indicating that the context account cannot predict meaningful effects of context reinstatement at test on the size of RIF. We found support for a null result (i.e., no differences in the size of RIF across groups) using a Bayesian odds analysis for ANOVA with odds 5.20 in favor of the null hypothesis (Masson, 2011). This null result is particularly problematic for the context account (i.e., both tenets acting in concert as the underlying proposed mechanism), because preferentially reinstating the context of learning at test should neutralize the RIF effect. If the context of learning is active at test, this account predicts that there should be no difference between Nrp[–] and Rp– items because all items were encountered in this initial phase. Our results did not demonstrate attenuated RIF and instead (nonsignificantly) tended in the opposite direction of the predicted effect.

We also observed no reliable effect of learning/retrieval-practice-context match, $F(1, 143) = 0.11$, $p = .745$, difference in *M*s = −0.02, 95% CI [−0.06, 0.04]. With 95% confidence in the direction predicted by the context account (i.e., more RIF with learning/retrieval-practice-context mismatch than match), the largest standardized mean difference that could be accounted for by the context account is 0.15, indicating that Tenet 1 of the context account can predict, at most, a small effect. A null result was moreover supported by a Bayesian odds analysis with odds 11.43 in favor of the null hypothesis, indicating that Tenet 1 of the context account fails to independently predict differences in the size of RIF using this preparation.

Retrieval-practice/test-context match did not interact with practice, indicating that the presence of a context shift between RP and test did not reliably affect the size of RIF, $F(1, 143) = 0.80$, $p = .372$, difference in *M*s = −0.05, 95% CI [−0.06, 0.03]. With 95% confidence in the direction predicted by the context account, the largest standardized mean difference that could be accounted for by the context account is 0.07, indicating that Tenet 2 of the context account cannot predict any meaningful differences in the size of RIF. This result was also supported by a Bayesian odds analysis with odds 5.90 in favor of the null hypothesis (Masson, 2011).

Although there were no observed differences between conditions, we wanted to be certain that RIF was reliable across conditions, so we compared the mean RIF score of each group with zero. We observed marginal RIF (indicated by lower recall of Rp– items than Nrp[–] items) in Condition AAA, $t(143) = 1.88$, $p = .062$, Cohen's $d = 0.31$, 95% CI [−0.01, 0.64], and in Condition AAB, $t(143) = 1.88$, $p = .063$, Cohen's $d = 0.31$, 95% CI [−0.01, 0.64]. We saw reliable RIF in Condition BAB, $t(143) = 2.54$, $p = .012$, Cohen's $d = 0.42$, 95% CI [0.09, 0.75], but no reliable RIF in Condition BAA, $t(143) = 0.39$, $p = .698$. These tendencies are opposite the predictions of the context account, which predicts no RIF effect in Condition BAB, especially relative to Condition BAA, in which the context account predicts the largest RIF effect (although this difference failed to reach significance, $t[143] = 1.56$, $p = .121$, difference in *M*s = −0.05, 95% CI [−0.10, 0.01]).

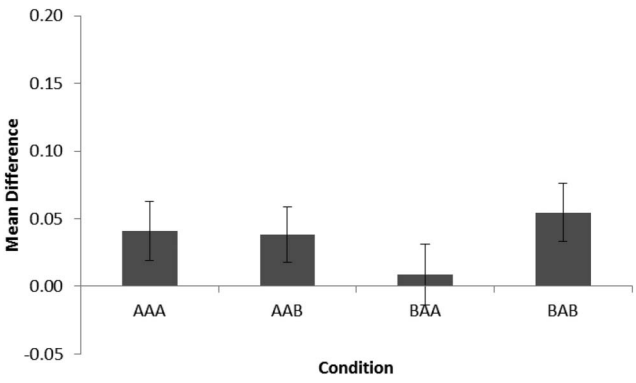

Figure 4. Mean difference scores reflecting the retrieval-induced forgetting in high-frequency (Rp– and Nrp[–]) items observed by condition in Experiment 2. These difference scores reflect the proportion of nontarget items in retrieval practiced categories (Rp–) relative to baseline items in comparable nonretrieval practiced categories (Nrp[–]; i.e., Nrp[–] – Rp–). As such, positive values indicate more retrieval-induced forgetting. The first letter indicates the context of learning, whereas the second letter signifies the context of retrieval practice, and the last letter indicates the context of test. Error bars indicate standard error of the means, but they should viewed cautiously, as the experiment was fully within subjects. Note that the vertical scale used here differs from that of Figure 2.

Nrp item assessment of context dependency. Nrp item performance was analyzed to assess the salience of our context shifts using a 2 (learning/test-context match) × 2 (learning/retrieval-practice-context match) × 2 (taxonomic frequency [+ vs. – items]) ANOVA. We observed no interactions between these factors (largest $p = .293$). However, we observed a strong effect of taxonomic word frequency, with high-frequency items better recalled than low-frequency items, $F(1, 144) = 403.21$, $p < .001$, Cohen's $d = 3.34$, 95% CI [2.84, 3.85]. We also found an effect of learning/retrieval-practice-context match, with items learned and Rp in the same contexts better recalled at test than items learned and practiced in different contexts, $F(1, 144) = 4.27$, $p = .040$, Cohen's $d = 0.34$, 95% CI [0.02, 0.67]. However, we found no evidence of a difference based on context shifts between learning and test context match, $F(1, 144) = 1.50$, $p = .215$ (Bayes odds 5.57 in favor of the null). This indicates that manipulations in presentation style were not sufficient to reliably create context shifts between learning and test.

Correlations. We failed to replicate Experiment 1's finding of a significant negative correlation overall between retrieval benefit and RIF, $r(142) = -0.05$, $p = .241$, 95% CI [−0.21, 0.12]. However, when we pooled the data across both experiments, this negative correlation reached significance, $r(264) = -0.08$, $p = .022$, 95% CI [−0.20, 0.04].

General Discussion

The present experiments tested predictions of the context account of RIF by manipulating contextual features of a typical retrieval practice paradigm. We view this test of the context account as appropriate because the account was designed to explain findings pertaining to a typical retrieval practice paradigm. Specifically, we manipulated contextual features similar (but admittedly not identical) to those posited to control RIF (i.e., differences in task presentation and demands during each phase of a typical RIF experiment), rather than the means through which Jonker et al. (2013) manipulated cognitive context (using videos and imagination tasks that are atypical of most RIF experiments). It is possible that these differences in manipulating context shifts can account for why our results did not replicate those of Jonker and colleagues. However, as our procedures were more consistent with the bulk of the RIF literature—the findings of which the context account attempts to speak to—and we manipulated context using features that proponents of this account argue create context shifts in a prototypical RIF paradigm, we think that that the present experiments were a fair test of the context account. Of course no single set of experiments ever definitively determines the validity of a hypothesis such as the context account of RIF, and future research should continue to consider situations in which the context account may contribute to RIF.

Taken together, the results of Experiments 1 and 2 failed to support the predictions of the context account. Experiment 1 tested Tenet 1 of the context account by minimizing the context shift between learning and retrieval practice in a generate condition. This change had no observable effect on the magnitude of RIF relative to a typical RIF paradigm with conventional study in Phase 1. Proponents of the context account could argue that the subtle change between generate during learning and retrieval practice (i.e., generate and generate again, as there were no additional instructions) was salient enough to constitute a context shift for participants. Experiment 2 refuted this possibility by failing to detect any observable effect of shifting more noticeable procedural features of the context between retrieval practice and test, specifically presentation style and spacing. As such, proponents of the context account would have difficulty arguing that salient features of the context, such as presentation style and spacing, do not constitute an appreciable context shift for participants, and at the same time argue that the difference between generation in Phase 1 and retrieval practice in Phase 2 was sufficient to provide a context shift. Thus, application of the context account to both experiments conjointly is untenable.

Experiment 2 tested both tenets (as well as the prediction they make conjointly) and also conceptually replicated the finding in Experiment 1 of reliable RIF following generative learning. None of the results from either experiment were consistent with the predictions made by the context account—neither each tenet independently, nor the prediction made by the underlying mechanism of both tenets working in concert.

Although these results are predicated on interpreting null findings, all of the critical results were supported by Bayesian odds analyses. In addition, the context account did not manage to predict any substantial standardized differences in RIF with 95% confidence in the predicted direction. That is, we can say with 95% confidence (i.e., $\alpha = .05$) that the context account fails to predict any appreciable standardized mean differences in the magnitude of the RIF effect in the study population using our manipulations. Both experiments also used large samples to obtain sufficiently narrow confidence intervals as well as provide meaningfully large Bayes odds in favor of the null. For these reasons, we feel that these results, though null, are supported by rigorous statistical tests.

The results of Experiment 1 and Experiment 2 are also problematic for interference accounts of RIF. These accounts posit that RIF depends on the strengthening of Rp+ items (i.e., the testing effect), and that this strengthening of practiced items is proportional to the forgetting of nonpracticed items as a result of retrieval or response competition between Rp+ and Rp– items at test (McGeoch, 1942). Other interference accounts propose some sort of associative blocking at test by strengthened items (Raaijmakers & Shiffrin, 1981). These types of accounts predict a positive correlation between the retrieval benefit for Rp+ items and RIF for Rp– items within subjects. Contrary to this prediction, we found a negative correlation between RIF and the testing effect collapsing across the two experiments. Our results are, for the most part, consistent with the literature investigating this correlation. Previous studies have observed trends toward negative correlations between retrieval benefit and RIF, but these tendencies often fail to reach statistical significance (e.g., Aslan & Bäuml, 2011; Hanslmayr, Staudigl, Aslan, & Bäuml, 2010; Staudigl, Hanslmayr, & Bäuml, 2010). This negative correlation is not only problematic for associative interference accounts of RIF but also for phenomena such as the list-strength effect, which has been accounted for as a strength-dependent memory outcome (Ratcliff, Clark, & Shiffrin, 1990), but can also be accounted for using an inhibitory mechanism (Bäuml, 1997).

Although these experiments were not designed to explicitly test inhibitory accounts, they are congruent with an inhibitory account of RIF (see Anderson, 2003; Anderson et al., 1994), which can

provide a post hoc explanation of our results. A general inhibitory account predicts no mediating effect of context shifts on RIF, consistent with what was observed in both Experiments 1 and 2. That is, because both experiments involved retrieval practice, the inhibitory account predicts RIF overall in both experiments, in all practiced categories (retrieval specificity). Furthermore, the active suppression proposed by inhibitory accounts theoretically does not depend on strengthening of Rp+ items. According to this account, although *attempting* retrieval is necessary to observe RIF, the success of this retrieval is not required (Storm et al., 2006). Hence, an inhibitory account does not require a positive correlation between the extent to which individuals benefit from retrieval practice and the extent to which they forget nontarget items. Although inhibitory accounts do not predict the observed negative correlation (rather, they predict *no* correlation), it is not problematic for inhibitory accounts to the same extent that it is for interference accounts. Therefore, the aforementioned negative correlation between retrieval benefit and RIF is less a problem for the inhibitory account.

These results speak not only to the Jonker and colleagues' (2013) context account of RIF, but by extension they also speak to context-based accounts of other retrieval phenomena, such as retrieval-based learning (Karpicke et al., 2014). The context account of RIF assumes reinstatement of either the study context or retrieval practice context during test as a function of which is most similar to the test context for the category immediately being tested. This account mirrors the reinstatement assumed to occur according to Karpicke et al.'s (2014) episodic context account of retrieval-based learning. They argue that the testing effect (i.e., the benefit of retrieval practice over additional study) occurs as a result of subjects' reinstatement of the context of study during retrieval practice (which saliently differs from study as a result of the procedure of each task and the constant change in context resulting from the passage of time), and a subsequent updating of the context associated with the practiced items. As such, items that were Rp should subsequently be readily retrievable at test once either the study context or the retrieval practice context has been reinstated.

When applied to the retrieval practice paradigm, the Karpicke et al. (2014) context account and the Jonker et al. (2013) context account appear to provide similar mechanisms and predictions. Although Jonker et al. emphasize procedural factors and Karpicke et al. focus more on the passage of time, both suggest that retrieval practice serves to update information in a second context, and produce the effects observed during test as a result of reinstatement of context cues facilitating recall of certain items. Karpicke et al.'s context account seems to accurately predict some the results from Experiment 1 (the finding of an attenuated retrieval benefit when the context shift between study and retrieval practice was weakened), but this does not carry through to the findings of Experiment 2. These findings suggest that a context account of testing effects depends on differences in procedure beyond presentational or environmental context shifts. Still, we did not have the proper baselines of comparison to speak effectively to the testing effect, which is typically examined with a restudied baseline.

According to these results, context shifts alone cannot wholly account for the RIF effect, particularly when the presentation style and task demands are used to manipulate context, consistent with the context shifts argued to produce a RIF effect in a standard retrieval practice paradigm. Although context shifts may contribute to RIF in certain circumstances, the present experiments did not lend any support to the notion that context shifts exclusively control the occurrence of RIF and provide a completely "inhibition-free" mechanism.

References

Anderson, M. C. (2003). Rethinking interference theory: Executive control and the mechanisms of forgetting. *Journal of Memory and Language, 49,* 415–445. http://dx.doi.org/10.1016/j.jml.2003.08.006

Anderson, M. C., & Bell, T. (2001). Forgetting our facts: The role of inhibitory processes in the loss of propositional knowledge. *Journal of Experimental Psychology: General, 130,* 544–570. http://dx.doi.org/10.1037/0096-3445.130.3.544

Anderson, M. C., Bjork, E. L., & Bjork, R. A. (2000). Retrieval-induced forgetting: Evidence for a recall-specific mechanism. *Psychonomic Bulletin & Review, 7,* 522–530. http://dx.doi.org/10.3758/BF03214366

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. http://dx.doi.org/10.1037/0278-7393.20.5.1063

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. http://dx.doi.org/10.1037/0033-295X.102.1.68

Aslan, A., & Bäuml, K. H. (2010). Retrieval-induced forgetting in young children. *Psychonomic Bulletin & Review, 17,* 704–709. http://dx.doi.org/10.3758/PBR.17.5.704

Aslan, A., & Bäuml, K. H. (2011). Individual differences in working memory capacity predict retrieval-induced forgetting. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 37,* 264–269. http://dx.doi.org/10.1037/a0021324

Bäuml, K. H. (1997). The list-strength effect: Strength-dependent competition or suppression? *Psychonomic Bulletin & Review, 4,* 260–264. http://dx.doi.org/10.3758/BF03209403

Bäuml, K. H. (2002). Semantic generation can cause episodic forgetting. *Psychological Science, 13,* 356–360. http://dx.doi.org/10.1111/j.0956-7976.2002.00464.x

Bjork, R. A. (1975). Retrieval as a memory modifier. In R. Solso (Ed.), *Information processing and cognition: The Loyola Symposium* (pp. 123–144). Hillsdale, NJ: Erlbaum.

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

Bower, G. H. (1981). Mood and memory. *American Psychologist, 36,* 129–148. http://dx.doi.org/10.1037/0003-066X.36.2.129

Camp, G., Pecher, D., & Schmidt, H. G. (2007). No retrieval-induced forgetting using item-specific independent cues: Evidence against a general inhibitory account. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 33,* 950–958. http://dx.doi.org/10.1037/0278-7393.33.5.950

Chan, J. C., Erdman, M. R., & Davis, S. D. (2015). Retrieval induces forgetting, but only when nontested items compete for retrieval: Implication for interference, inhibition, and context reinstatement. *Journal of Experimental Psychology: Learning, Memory, and Cognition*. Advance online publication. http://dx.doi.org/10.1037/xlm0000101

Ciranni, M. A., & Shimamura, A. P. (1999). Retrieval-induced forgetting in episodic memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 25,* 1403–1414. http://dx.doi.org/10.1037/0278-7393.25.6.1403

Dulsky, S. G. (1935). The effect of a change of background on recall and relearning. *Journal of Experimental Psychology, 18,* 725–740. http://dx.doi.org/10.1037/h0058066

Eich, E., & Metcalfe, J. (1989). Mood dependent memory for internal versus external events. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 15,* 443–455. http://dx.doi.org/10.1037/0278-7393.15.3.443

Eich, J. E., Weingartner, H., Stillman, R. C., & Gillin, J. C. (1975). State-dependent accessibility of retrieval cues in the retention of a categorized list. *Journal of Verbal Learning & Verbal Behavior, 14,* 408–417. http://dx.doi.org/10.1016/S0022-5371(75)80020-X

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. http://dx.doi.org/10.1111/j.2044-8295.1975.tb01468.x

Goodwin, D. W., Powell, B., Bremer, D., Hoine, H., & Stern, J. (1969). Alcohol and recall: State-dependent effects in man. *Science, 163,* 1358–1360. http://dx.doi.org/10.1126/science.163.3873.1358

Grundgeiger, T. (2014). Noncompetitive retrieval practice causes retrieval-induced forgetting in cued recall but not in recognition. *Memory & Cognition, 42,* 400–408. http://dx.doi.org/10.3758/s13421-013-0372-z

Hanslmayr, S., Staudigl, T., Aslan, A., & Bäuml, K. H. (2010). Theta oscillations predict the detrimental effects of memory retrieval. *Cognitive, Affective & Behavioral Neuroscience, 10,* 329–338. http://dx.doi.org/10.3758/CABN.10.3.329

Hicks, J. L., & Starns, J. J. (2004). Retrieval-induced forgetting occurs in tests of item recognition. *Psychonomic Bulletin & Review, 11,* 125–130. http://dx.doi.org/10.3758/BF03206471

Jakab, E., & Raaijmakers, J. G. (2009). The role of item strength in retrieval-induced forgetting. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 35,* 607–617. http://dx.doi.org/10.1037/a0015264

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. http://dx.doi.org/10.1037/0278-7393.34.1.112

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. http://dx.doi.org/10.1037/a0027277

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

Karpicke, J. D., Lehman, M., & Aue, W. R. (2014). Retrieval-based learning: An episodic context account. *Psychology of Learning and Motivation, 61,* 237–284. http://dx.doi.org/10.1016/B978-0-12-800283-4.00007-1

Kuhl, B. A., Dudukovic, N. M., Kahn, I., & Wagner, A. D. (2007). Decreased demands on cognitive control reveal the neural processing benefits of forgetting. *Nature Neuroscience, 10,* 908–914. http://dx.doi.org/10.1038/nn1918

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

McGeoch, J. A. (1942). *The psychology of human learning*. New York, NY: Longmans & Green.

Miguez, G., Mash, L. E., Polack, C. W., & Miller, R. R. (2014). Failure to observe renewal following retrieval-induced forgetting. *Behavioural Processes, 103,* 43–51. http://dx.doi.org/10.1016/j.beproc.2013.11.008

Morris, C. D., Bransford, J. D., & Franks, J. J. (1977). Levels of processing versus transfer appropriate processing. *Journal of Verbal Learning & Verbal Behavior, 16,* 519–533. http://dx.doi.org/10.1016/S0022-5371(77)80016-9

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,* 1383–1409. http://dx.doi.org/10.1037/a0037505

Ortega-Castro, N., & Vadillo, M. A. (2013). Retrieval-induced forgetting and interference between cues: Training a cue-outcome association attenuates retrieval by alternative cues. *Behavioural Processes, 94,* 19–25. http://dx.doi.org/10.1016/j.beproc.2012.11.010

Overton, D. A. (1971). Discriminative control of behavior by drug states. In T. Thompson & R. Pickens (Eds.), *Stimulus properties of drugs* (pp. 87–110). New York, NY: Appleton-Century-Crofts.

Raaijmakers, J. G. W., & Jakab, E. (2012). Retrieval-induced forgetting without competition: Testing the retrieval specificity assumption of the inhibition theory. *Memory & Cognition, 40,* 19–27. http://dx.doi.org/10.3758/s13421-011-0131-y

Raaijmakers, J. G., & Jakab, E. (2013). Rethinking inhibition theory: On the problematic status of the inhibition theory for forgetting. *Journal of Memory and Language, 68,* 98–122. http://dx.doi.org/10.1016/j.jml.2012.10.002

Raaijmakers, J. G. W., & Shiffrin, R. M. (1981). Search of associative memory. *Psychological Review, 88,* 93–134. http://dx.doi.org/10.1037/0033-295X.88.2.93

Ratcliff, R., Clark, S. E., & Shiffrin, R. M. (1990). List-strength effect: I. Data and discussion. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 16,* 163–178. http://dx.doi.org/10.1037/0278-7393.16.2.163

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

Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., & Iverson, G. (2009). Bayesian t-tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review, 16,* 225–237. http://dx.doi.org/10.3758/PBR.16.2.225

Sahakyan, L., & Hendricks, H. E. (2012). Context change and retrieval difficulty in the list-before-last paradigm. *Memory & Cognition, 40,* 844–860. http://dx.doi.org/10.3758/s13421-012-0198-0

Smith, S. M. (1979). Remembering in and out of context. *Journal of Experimental Psychology: Human Learning and Memory, 5,* 460–471. http://dx.doi.org/10.1037/0278-7393.5.5.460

Smith, S. M., Glenberg, A., & Bjork, R. A. (1978). Environmental context and human memory. *Memory & Cognition, 6,* 342–353. http://dx.doi.org/10.3758/BF03197465

Staudigl, T., Hanslmayr, S., & Bäuml, K. H. T. (2010). Theta oscillations reflect the dynamics of interference in episodic memory retrieval. *The Journal of Neuroscience, 30,* 11356–11362. http://dx.doi.org/10.1523/JNEUROSCI.0637-10.2010

Storm, B. C., Angello, G., Buchli, D. R., Koppel, R. H., Little, J. L., & Nestojko, J. F. (2015). A review of retrieval-induced forgetting in the contexts of learning, eyewitness memory, social cognition, autobiographical memory, and creative cognition. *Psychology of Learning and Motivation, 62,* 141–194. http://dx.doi.org/10.1016/bs.plm.2014.09.005

Storm, B. C., Bjork, E. L., & Bjork, R. A. (2008). Accelerated relearning after retrieval-induced forgetting: The benefit of being forgotten. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 34,* 230–236. http://dx.doi.org/10.1037/0278-7393.34.1.230

Storm, B. C., Bjork, E. L., Bjork, R. A., & Nestojko, J. F. (2006). Is retrieval success a necessary condition for retrieval-induced forgetting? *Psychonomic Bulletin & Review, 13,* 1023–1027. http://dx.doi.org/10.3758/BF03213919

Storm, B. C., Friedman, M. C., Murayama, K., & Bjork, R. A. (2014). On the transfer of prior tests or study events to subsequent study. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 40,* 115–124. http://dx.doi.org/10.1037/a0034252

Storm, B. C., & Levy, B. J. (2012). A progress report on the inhibitory account of retrieval-induced forgetting. *Memory & Cognition, 40,* 827–843. http://dx.doi.org/10.3758/s13421-012-0211-7

Tulving, E., & Thomson, D. M. (1973). Encoding specificity and retrieval processes in episodic memory. *Psychological Review, 80,* 352–373. http://dx.doi.org/10.1037/h0020071

Van Overschelde, J. P., Rawson, K. A., & Dunlosky, J. (2004). Category norms: An updated and expanded version of the Battig and Montague (1969) norms. *Journal of Memory and Language, 50,* 289–335. http://dx.doi.org/10.1016/j.jml.2003.10.003

Veling, H., & van Knippenberg, A. (2004). Remembering can cause inhibition: Retrieval-induced inhibition as cue independent process. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 30,* 315–318. http://dx.doi.org/10.1037/0278-7393.30.2.315

Verde, M. F. (2012). 2 Retrieval-induced forgetting and inhibition: A critical review. *Psychology of Learning and Motivation, 56,* 47–80. http://dx.doi.org/10.1016/B978-0-12-394393-4.00002-9

Williams, C. C., & Zacks, R. T. (2001). Is retrieval-induced forgetting an inhibitory process? *The American Journal of Psychology, 114,* 329–354. http://dx.doi.org/10.2307/1423685

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. http://dx.doi.org/10.1038/nn.3973

Wimber, M., Rutschmann, R. M., Greenlee, M. W., & Bäuml, K. H. (2009). Retrieval from episodic memory: Neural mechanisms of interference resolution. *Journal of Cognitive Neuroscience, 21,* 538–549. http://dx.doi.org/10.1162/jocn.2009.21043

Received February 26, 2015
Revision received June 11, 2015
Accepted June 15, 2015 ■