Journal of Memory and Language 102 (2018) 83–96

Contents lists available at ScienceDirect

Journal of Memory and Language

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

Testing potentiates new learning across a retention interval and a lag: A strategy change perspective

Jason C.K. Chan[a,*], Krista D. Manley[a], Sara D. Davis[a], Karl K. Szpunar[b]

[a] Iowa State University, USA
[b] University of Illinois at Chicago, USA

ARTICLE INFO

Keywords:
Retrieval practice
New learning
Test-potentiated learning
Forward testing effect
Relational processing
Strategy change

ABSTRACT

Practicing retrieval on previously studied materials can potentiate subsequent learning of new materials. In four experiments, we investigated the influence of retention interval and lag on this test-potentiated new learning (TPNL) effect. Participants studied four word lists and either practiced retrieval, restudied, or completed math problems following Lists 1–3. Memory performance on List 4 provided an estimate of new learning. In Experiments 1 and 2, participants were tested on List 4 after either a 1 min or 25 min retention interval. In Experiments 3 and 4, participants took at 25 min break before studying List 4. A TPNL effect was observed in all experiments. To gain insight into the mechanism that may underlie TPNL, we analyzed the extent to which participants organized their recall from list to list. Relative to restudy and math, testing led to superior semantic organization across lists. Our results support a strategy change account of TPNL.

Introduction

A growing body of research has shown that interspersing encoding with test questions can strengthen student learning. For example, when viewing a lecture, students who answer quiz questions throughout a lecture often better remember the *tested* information than students who are not quizzed (i.e., the testing effect, McDaniel, Roediger, & McDermott, 2007). More important for present purposes, however, students who answer interspersed quiz questions also better learn new information presented *after* the quiz than students who are not quizzed (e.g., Jing, Szpunar, & Schacter, 2016; Szpunar, Khan, & Schacter, 2013). That is, interspersed testing enhances *new learning*. In this paper, we refer to this benefit of testing as test-potentiated new learning, or *TPNL*.

The TPNL effect is typically investigated using a multi-list or multi-section learning paradigm. For example, subjects may be asked to memorize two lists of words. After studying List 1, subjects may take a test for that list (i.e., the *interspersed-testing condition*) or not (i.e., the *no-testing condition*) before they study List 2. In this example, List 1 represents *original learning* and List 2 represents *new learning*, and differences in performance for List 2 between the interspersed-testing condition and the no-testing condition demonstrate the influence of testing on new learning. A wealth of research has shown that testing can facilitate learning of new information (Chan, Meissner, & Davis, in preparation; Pastotter & Bauml, 2014; Yang, Potts, & Shanks, 2018). In general, the TPNL effect is robust and applicable to a variety of learning situations. For example, testing can promote new learning of lists of single words (Szpunar, McDermott, & Roediger, 2008), word pairs (Tulving & Watkins, 1974; Wahlheim, 2015), picture-word pairs (Davis & Chan, 2015; Weinstein, McDermott, & Szpunar, 2011), text passages (Wissman & Rawson, 2015; Wissman, Rawson, & Pyc, 2011), and video lectures (Szpunar et al., 2013). However, several factors have also been shown to moderate this effect, including the participants' perceived likelihood of being tested (Weinstein, Gilmore, Szpunar, & McDermott, 2014; but see also Wissman et al., 2011) and the frequency with which participants have to switch between retrieval and new learning (Davis & Chan, 2015; Davis, Chan, & Wilford, 2017).

Despite the increasingly sizable literature on the phenomenon of TPNL, we currently know very little about the persistence of this effect across a time delay. In the present study, we focus on two kinds of time delay: (1) the *retention interval* between new learning and its assessment and (2) the time between presentation of original and new learning, which we refer to as "*lag.*" The dearth of research on delay is particularly glaring given the copious amount of evidence for the longevity of the testing effect (for reviews, see Adesope, Trevisan, & Sundararajan, 2017; Rowland, 2014). We now describe the importance of these two types of delay for both the application of interspersed testing to education and the theoretical understanding of TPNL.

* Corresponding author at: Department of Psychology, Iowa State University, Ames, IA 50011, USA.
E-mail address: ckchan@iastate.edu (J.C.K. Chan).

https://doi.org/10.1016/j.jml.2018.05.007
Received 5 October 2017; Received in revised form 12 May 2018
Available online 30 May 2018
0749-596X/ Published by Elsevier Inc.

Retention interval

To date, most studies in this literature have assessed the influence of interspersed testing on new learning at very short (1 min) retention intervals (Aslan & Bauml, 2015; Chan, Thomas, & Bulevich, 2009; Davis & Chan, 2015; Szpunar et al., 2008; Tulving & Watkins, 1974; Wahlheim, 2015; Weinstein et al., 2011; Wissman & Rawson, 2015), and the effects of retention interval on TPNL have rarely been the focus of extant investigations. Moreover, studies that have included multiple retention intervals have produced mixed results. For example, Szpunar et al. (2008) had participants study five lists of words, and participants either completed an immediate test or math problems following each of the first four lists. After studying a fifth list, all participants received a test for that list. The List 5 test allowed the researchers to examine the immediate impact of interspersed testing (relative to math) on new learning (of List 5). Furthermore, all participants took a final recall test of all studied items (including items from List 5) 30 min later, and the results of this delayed test showed that the TPNL effect persisted across the 30 min retention interval (see also Jing et al., 2016; Nunes & Weinstein, 2012; Pierce, Gallo, & McCain, 2017; Szpunar et al., 2013; Weinstein et al., 2011 in which the TPNL effect persisted across a retention interval of 5 min or less). In fact, across two experiments, the TPNL effect was nearly identical regardless of whether one examines performance in the 1-min (d = 1.52) or 30-min delay test (d = 1.60). In contrast, in an experiment that employed a similar design, Wissman and Rawson (2015 Experiment 4) found that the TPNL effect at immediate testing (d = 1.66) was substantially diminished after just a 15-min delay (d = 0.78), which suggests that the TPNL effect might be somewhat ephemeral. Indeed, a recent meta-analysis found that studies that used longer retention intervals tended to produce a smaller TPNL effect than studies that employed shorter retention intervals (Chan et al., in preparation). However, as with all moderator analyses that include data from different studies, the result of this meta-regression is correlational in nature and must be interpreted with caution. Hence, more research is needed to evaluate the influence of retention interval on TPNL, particularly if the benefits of interpolated testing are to be interpreted as having relevance for educational practice.

Perhaps even more important than the mixed results on TPNL and retention interval is that interpretation of final test performance in existing studies is not straightforward. For instance, Szpunar et al. (2008) and Wissman and Rawson (2015) required participants to take *both an immediate test and a final test* for the new learning materials. Therefore, recall performance on the delayed final test was contaminated by that of the immediate test, making it difficult to estimate the true effects of retention interval on TPNL. In Experiments 1 and 2, we sought to assess the influence of retention interval on TPNL without this potential source of contamination. Specifically, we administered only one test for the critical new learning material after a filled retention interval of either 1 min (to clear short-term memory) or 25 min.

Lag

To our knowledge, no studies to date have examined the influence of lag (i.e., the delay between original learning and new learning) on TPNL in a multi-list learning paradigm. The effects of lag on TPNL have important implications both for the implementation of interspersed testing in the classroom and the theoretical understanding of TPNL. Specifically, pedagogical guides often stress the fact that sustaining attention for long periods of time can be difficult, which can lead to frequent mind wanderings by the learner (Bunce, Flens, & Neiles, 2010; Risko, Anderson, Sarwal, Engelhardt, & Kingstone, 2012; Szpunar, 2017). Educators possess an intuitive understanding of the fact that learners often struggle to sustain attention, and they suggest taking breaks as a strategy for counteracting the negative impact of time-on-task for learning information presented at the end of long study sequences. These breaks can take various forms, including asking learners questions (i.e., testing), presenting a video, having group discussions, or simply giving students a bathroom break (Centre for Teaching Excellence - University of Waterloo., 2012; Olmsted, 1999). Such study breaks are thought to be effective at helping students refocus and learn new information, because they allow students to temporarily deactivate the prolonged task-goal of learning and attend to activities with different task goals (Ariga & Lleras, 2011). Similarly, taking interspersed tests can enhance new learning by providing a break from the encoding activities required by a prolonged study sequence. Specifically, forcing participants to switch the task from encoding to retrieval has been hypothesized to initiate a context change (Jang & Huber, 2008; Jonker, Seli, & MacLeod, 2013; Whiffen & Karpicke, 2017), which allows participants to "reset" their encoding operations (Pastotter, Schicker, Niedernhuber, & Bauml, 2011).

An important question is whether testing enhances new learning because it essentially serves as a study break, or if retrieval is "special" in its ability to enhance new learning beyond providing a break to encoding activities. If the former possibility proves correct, then testing should not facilitate new learning when compared to a condition in which new learning occurs following a study break. In Experiments 3 and 4, we aimed to examine the influence of lag on TPNL. Specifically, participants studied four lists of words. After presentation of each of the first three lists, participants either recalled the list, performed mental arithmetic, or restudied the list. To examine whether the benefits of testing on new learning are distinct from those afforded by providing a study break, we inserted a 25-min filled lag just before participants studied List 4. During this lag, participants took a break from encoding by completing a series of brain teasers and then playing the videogame Tetris (more details about these tasks are described in the Method section of Experiment 3). These tasks were selected as the lag activities because they differed substantially from the encoding task.

A strategy change perspective of test-potentiated new learning

In the preceding section, we described one potential mechanism by which interpolating retrieval can facilitate subsequent learning – namely, that changing from an encoding context to a retrieval context may provide a break from the encoding activities (Jang & Huber, 2008). Research in verbal learning (Gunter, 1980; Wickens, 1970) and intentional forgetting (Sahakyan & Kelley, 2002) have repeatedly demonstrated that changing context can release learners from the negative impact of proactive interference, thereby facilitating new learning. From this perspective, inserting memory tests and inserting study breaks into an encoding session may serve similar functions. An alternative account, however, posits that interpolated testing enhances new learning beyond context change — specifically, testing may enhance new learning by offering an opportunity for participants to switch to more effective encoding strategies during new learning.

According to this strategy change account, taking a memory test can lead participants to use different, and perhaps superior, encoding strategies for later learning (Cho, Neely, Crocco, & Vitrano, 2017; Gordon & Thomas, 2017), because the test provides participants with important, performance-relevant information such as test format, the type of retrieval cues available, the amount of time available for retrieval, etc. The idea that performing retrieval can alter how participants approach subsequent learning has received some empirical support. For example, learners reported that they were more likely to use deeper encoding strategies when relearning previously studied materials after a test trial than after a restudy trial (Soderstrom & Bjork, 2014). Further, taking a test can alter how participants distribute their encoding or attentional resources during subsequent encoding opportunities (Chan, Manley, & Lang, 2017; Gordon & Thomas, 2014; Jing et al., 2016; Szpunar et al., 2013). For example, in a recent study using a triad learning paradigm (Davis & Chan, 2015; see also Finn & Roediger, 2013), participants first studied a set of face-name pairs. Next, participants either restudied or recalled the name associated with

each face (i.e., original learning) before they studied the *profession* for that face (i.e., new learning). Importantly, during this new learning trial, both the face-name (i.e., original learning) and face-profession (i.e., new learning) associations were present for study. Surprisingly, instead of demonstrating the usual TPNL effect, testing impaired learning of the new, face-profession association. Through several experiments, Davis and Chan (2015) attributed this result to the fact that attempting to retrieve the face-name association altered how participants approached the encoding task when the face-profession association (along with the face-name association) was presented for encoding. Specifically, they argued that the face-name association test trial, but not the restudy trial, revealed to participants the difficulty of learning the face-name pair. When the new, face-profession association was presented for study, participants "borrowed time" from the new-learning trial to restudy the face-name association, thus impairing new learning. Moreover, recent research has shown that testing can affect both test expectancy and the amount of time participants spend on future learning activities. For example, taking a test increases learners' expectation that they will be tested again in the near future (Weinstein et al., 2014). Perhaps partly because of this increased test expectancy, when learners were allowed to self-regulate their study duration, those who received interpolated tests spent longer to study new information than those who did not (Gordon & Thomas, 2014; Gordon, Thomas, & Bulevich, 2015; Yang, Potts, & Shanks, 2017).

Although the results of the above-cited studies are consistent with the idea that retrieval can cause a strategy change for subsequent encoding, they do not provide direct evidence that strategy change underlies TPNL. Specifically, Soderstrom and Bjork's (2014) results were based on a relearning, not new-learning, paradigm, and the data regarding strategy change were based on participants' subjective report. Moreover, Davis and Chan (2015) did not provide direct evidence of strategy change, because they did not measure the amount of time that participants devoted to relearning of the face-name association relative to new learning of the face-profession association. Lastly, although the findings that prior testing increases test expectancy (Weinstein et al., 2014) and new-learning duration (Gordon et al., 2015; Yang et al., 2017) signal a shift in strategy, these findings do not provide evidence that the strategy change is qualitative in nature. In the present experiments, we attempted to provide a more direct test of this strategy change account by examining organization in recall on a list-by-list basis.

Prior work on interpolated testing has focused primarily on the quantity of the learning that takes place during new learning trials (e.g., how many words are correctly recalled), and little is known about the quality of the learning. To the extent that the type of strategy that participants use to encode items is reflected in the way they recall these items, we can assess their strategy use by examining how they organize their recall. Indeed, prior work has shown that interpolated testing can serve to boost integration of information presented within and across video lecture segments *on a final cumulative test* (Jing et al., 2016). Nonetheless, no such analyses have been conducted in the context of TPNL experiments using word list stimuli, and more importantly, in a manner that assesses response organization *during initial and new learning*. Critically, a list-by-list analysis of output order based on semantic clustering is necessary to understand how interpolated testing affects participants' approach to retrieval, and perhaps encoding, of the lists. To address this gap in the literature, we asked participants in all four experiments to learn lists comprising words that belonged to several categories, and we analyzed the extent to which free recall of each list was characterized by category-based clustering. Recent work has shown that testing can serve to boost category clustering of word stimuli (Zaromb & Roediger, 2010). Accordingly, we predicted that interpolated testing should result in higher levels of category clustering during new learning as compared to no-testing and restudying.

To summarize, in the present experiments, we sought to examine whether time delay alters the beneficial effects of testing on new learning. Specifically, we compared testing with no-testing in Experiment 1, and we compared testing with restudying in Experiment 2. In each of these experiments, we examined the magnitude of the TPNL effect following a 1-min or 25-min retention interval. In Experiments 3 and 4, we examined the effects of lag on TPNL. Here, new learning occurred following either a 1-min lag (in Experiment 4) or a 25-min lag (in Experiments 3 and 4), and we compared testing to restudying and no-testing.

Experiment 1

Method

Design and participants

Intervening task (testing vs. no-testing) and retention interval (1 min vs. 25 min) were manipulated between-subjects. Participants were 186 undergraduate students from Iowa State University, who completed the experiment for course credit. English was not the primary language for 18 participants and their data were removed from analysis. Moreover, data from an additional 22 participants were removed because the experimenter ran the incorrect experiment program for the study phase and the delayed test. Therefore, data from 146 participants were analyzed. There were 36 participants in the *no-testing*, *1*-min *retention interval* condition, 39 participants in the *testing*, *1*-min *retention interval* condition, 37 participants in the *no-testing*, *25*-min *retention interval* condition, and 34 participants in the *testing*, *25*-min *retention interval* condition. We determined the desired sample size based on a *meta*-analytic effect size of TPNL ($g = 0.75$, Chan et al., in preparation). To achieve 85% power, each between-subjects condition required 34 participants.

Materials and procedure

Four interrelated lists with 15 words each were constructed. Each list contained three exemplars from five categories (Van Overschelde, Rawson, & Dunlosky, 2004). The five categories were animals, weather, fruits, human body parts, and building parts. Although the average taxonomic frequencies differed across the five categories ($M_{animals} = .17$, $M_{weather} = .15$, $M_{fruits} = .25$, $M_{bodyparts} = .33$, $M_{building} = .22$), $F(4, 55) = 4.63$, $p = .003$, they did not differ across the four lists (range = .21–.24), $F(3, 56) = 0.26$, $p = .86$, $B_{01} = 8.58$.

Fig. 1 illustrates the experimental design. Participants were informed that they would see several word lists of 15 words, with each word presented twice within a list.[1] They were also told that they would complete some math problems after studying each list, and then they would either take a memory test for the list or not, with the occurrence of the test being determined randomly by the computer. In actuality, participants were either tested after every list (testing condition) or only after List 4 (no-testing condition). In all memory tests, participants were told to recall words from only the most recent list, but all participants were told to expect a cumulative final test for all studied words.

For each list, a prompt (e.g., "This is Word List 1") appeared for 2 s, followed by a fixation cross that appeared in the middle of the screen for 1 s. Next, the words were presented for 4 s each, with the presentation of each word separated by a 500 ms blank interval. Each list was presented twice with no breaks in between, but a different random order was used during each presentation. List order was counterbalanced across participants. After studying each list, participants completed 60 s of math problems. Next, participants either completed an additional 60 s of math problems (no-testing condition) or they were given a *free recall* test for 60 s (testing condition).

In the 1-min retention interval conditions, the List 4 test began after

[1] We opted to present each word list twice because pilot testing ($N = 14$) revealed near-floor recall performance following a 25-min retention interval when the words were presented only once.

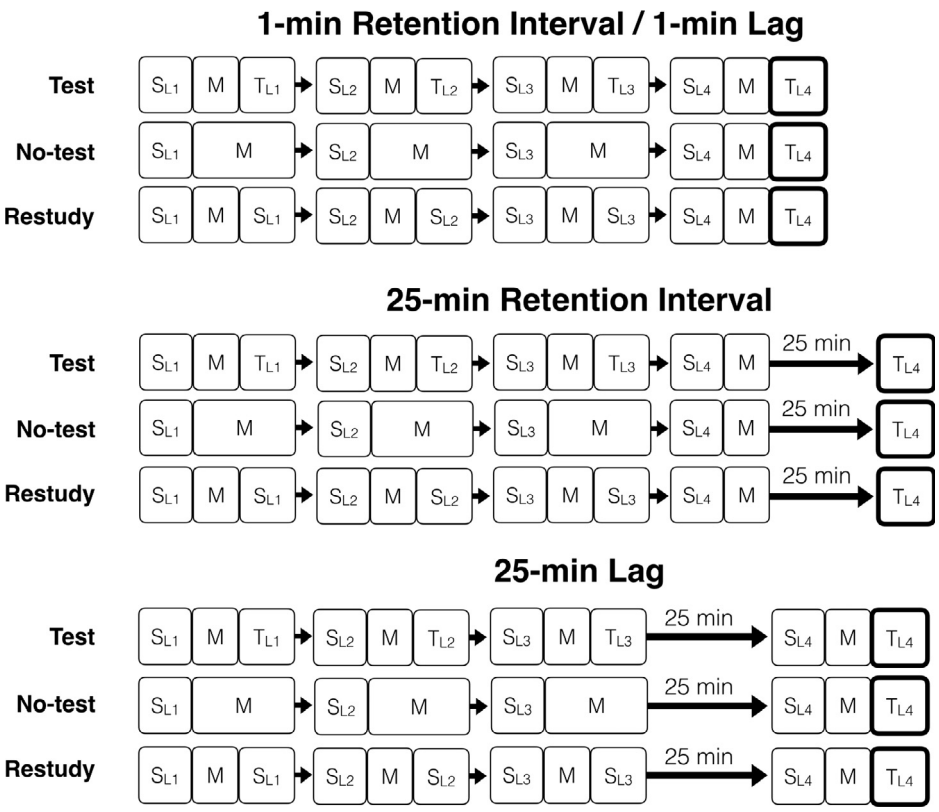


Fig. 1. Experimental design for the three experiments. Experiment 1 compared testing with no-testing across the 1-min and 25-min retention intervals. Experiment 2 compared testing with restudying across the 1-min and 25-min retention intervals. Experiment 3 compared testing with no-testing and restudying at a 25-min lag. Experiment 4 compared testing with restudying and no-testing across the 1-min and 25-min lags. S refers to study (or restudy), M refers to math problems and this phase lasted 1 min (hence the 1-min retention interval at the top of the figure), T refers to an interpolated free recall test, and L1-L4 in subscripts refer to Lists 1–4, respectively.

participants completed 1 min of math problems, and this applied to participants in both the no-testing and the testing conditions. The List 4 test was administered in the same fashion as the interspersed tests. That is, participants were instructed to recall as many words as possible from List 4 in 60 s.

In the 25-min retention interval conditions, participants completed the List 4 test following 25 min of brain teasers and the videogame Tetris. The brain teasers were displayed on the computer screen using a PowerPoint presentation and participants wrote their answers on paper. The brain teaser task contained 12 questions designed to assess abstract thinking and problem-solving skills (see the Appendix for examples). If participants finished the brain teasers within 25 min, they played the videogame Tetris for the remaining time. Tetris requires the arrangement of cascading blocks into complete lines, which are then cleared from the grid. Successful performance in Tetris likely requires effective coordination between spatial imagery (e.g., mental rotation) and motor skills.

Participants completed a source recognition test as the final task of the experiment. On each trial, participants saw a studied word and indicated its list membership (List 1, 2, 3, or 4) by pressing the corresponding number key; they then rated their confidence on a scale from 1 (very unsure) to 8 (very sure). Because performance on this source recognition test was necessarily contaminated by that of the recall tests, we opted to present data from the source test in the supplementary material and those data will not be discussed further.

Results and discussion

For all experiments, we first report results regarding the impact of interspersed testing and retention interval on correct recall, semantic clustering, and intrusions during the List 4 test. We then report recall performance across lists for participants in the testing condition (who were the only participants tested for the first three lists). Bayes factors (B_{01}, which indicates support for the null hypothesis over the alternative hypothesis) were provided when the result did not meet conventional level of statistical significance (i.e., $\alpha = .05$).

List 4 recall

Correct recall. We conducted a 2 (intervening task: testing vs. no-testing) × 2 (retention interval: 1 min vs. 25 min) between-subjects ANOVA to examine the effects of interspersed testing and retention interval on new learning (see the left side of Fig. 2). The dependent variable in this ANOVA was the proportion of List 4 words correctly recalled. The ANOVA revealed a main effect of intervening task, $F(1, 142) = 31.55$, $p < .01$, $\eta_p^2 = .15$. That is, participants who were tested on Lists 1–3 exhibited greater recall of List 4 ($M = .56$) than participants who were not tested ($M = .33$). The main effect of retention interval was also significant, $F(1, 142) = 36.13$, $p < .01$, $\eta_p^2 = .20$, with participants recalling fewer List 4 words after the 25-min retention interval ($M = .32$) than the 1-min retention interval ($M = .56$). Perhaps most important for present purposes, the interaction between intervening task and retention interval was not significant, $F(1, 142) = 1.09$, $p = .30$, $\eta_p^2 < .01$, $B_{01} = 4.81$, with the Bayes factor indicating that the data were nearly five times more probable under the null hypothesis than under the alternative hypothesis. This finding suggests that the beneficial effects of testing on new learning were observed at both the 1-min and 25-min retention intervals.

Clustering in recall. To investigate how interspersed testing influenced participants' use of strategies, we examined the likelihood with which participants clustered related items together during recall. As stated in the Method section, we spread words that belong to the same category across four lists and randomized the presentation order within a list. Consequently, words from the same category were often not presented on consecutive encoding trials. Previous research has shown that testing can improve semantic organization of studied material (Jing et al., 2016; Zaromb & Roediger, 2010). If this were the case in the

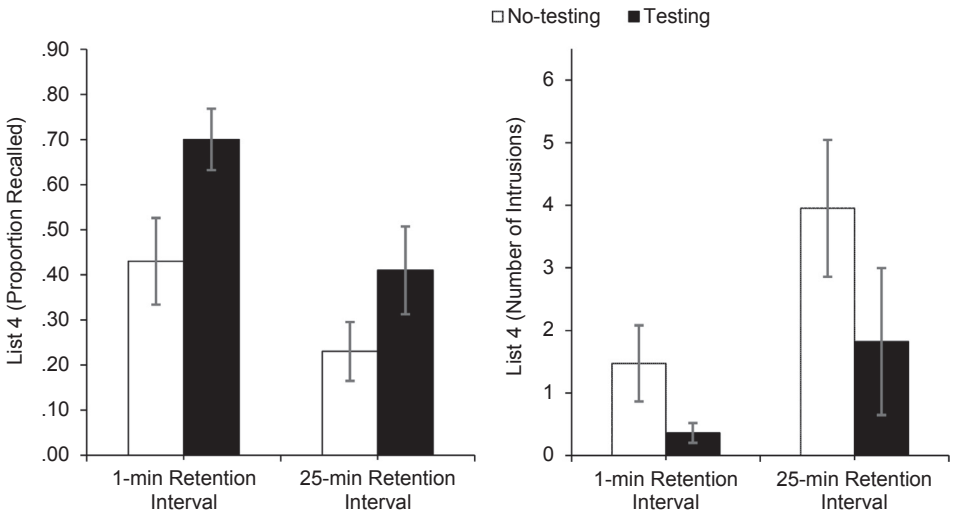


Fig. 2. Correct List 4 recall and intrusions as a function of intervening task and retention interval in Experiment 1. Left panel shows proportion of correct recall; right panel shows number of intrusions during List 4 recall. Error bars indicate descriptive 95% confidence intervals.

present context, testing should increase the clustering of related items during recall. Adjusted-ratio-of-clustering (ARC, Roenker, Thompson, & Brown, 1971) quantifies the likelihood that related items follow each other during output (i.e., clustering in recall), with positive ARC scores indicating above chance clustering, 0 indicating chance level clustering, and negative scores indicating below chance clustering. In this analysis, we substituted an undefined ARC score with 0, which occurs when only one item is recalled from each category or when all of the recalled items are from the same category.

A 2 (testing vs. no-testing) × 2 (1 min vs. 25 min) ANOVA showed a significant main effect for intervening task on *List 4 ARC* scores, $F(1, 142) = 19.19$, $p < .01$, $\eta_p^2 = 0.11$. The rightmost column in Table 1 depicts results of this analysis. Specifically, the tested participants clustered their output to a much greater degree ($M = .55$) than the nontested participants ($M = .20$). Retention interval also had an effect, $F(1, 142) = 4.54$, $p = .04$, $\eta_p^2 = .03$, with participants clustering less at the 25-min retention interval ($M = 0.29$) than at the 1-min retention interval ($M = .46$). The interaction, however, was not significant, $F(1, 142) = 2.22$, $p = .14$, $\eta_p^2 = .02$, $B_{01} = 1.66$. In sum, similar to the correct recall data, the clustering data showed that the benefits of interpolated testing on new learning persisted across the retention interval.

Table 1
ARC (Clustering) scores for experiments 1–4.

	List 1	List 2	List 3	List 4
Experiment 1				
1-min RI				
No-testing				0.34 (0.50)
Testing	0.32 (0.43)	0.59 (0.37)	0.57 (0.36)	0.58 (0.48)
25-min RI				
No-testing				0.05 (0.52)
Testing	0.48 (0.33)	0.41 (0.46)	0.65 (0.40)	0.52 (0.43)
Experiment 2				
1-min RI				
Restudying				0.20 (0.54)
Testing	0.29 (0.53)	0.56 (0.54)	0.58 (0.49)	0.57 (0.72)
25-min RI				
Restudying				0.19 (0.46)
Testing	0.30 (0.35)	0.53 (0.29)	0.57 (0.43)	0.49 (0.79)
Experiment 3				
25-min Lag				
No-testing				0.36 (0.58)
Restudying				0.31 (0.52)
Testing	0.35 (0.48)	0.61 (0.45)	0.62 (0.43)	0.58 (0.50)
Experiment 4				
1-min Lag				
No-testing				0.27 (0.59)
Restudying				0.18 (0.76)
Testing	0.29 (0.51)	0.58 (0.45)	0.55 (0.46)	0.71 (0.38)
25-min Lag				
No-testing				0.18 (0.58)
Restudying				0.22 (0.79)
Testing	0.37 (0.48)	0.52 (0.45)	0.66 (0.40)	0.59 (0.44)

Note: Standard deviations are in parentheses. Note that both the retention interval (RI) and lag manipulations did not occur until List 4.

Intrusions. In our experiments, participants were always told to recall words from the just-studied list. Therefore, when they recalled words from other lists, these items were considered intrusions. To examine the frequency with which intrusions occurred during the List 4 test, we conducted a 2 (testing vs. no-testing) × 2 (1 min vs. 25 min) ANOVA with the number of intrusions as the dependent variable. The means for this analysis are depicted in the right side of Fig. 2. The main effect of testing was significant, $F(1, 142) = 15.33$, $p < .01$, $\eta_p^2 = .10$, such that intrusions occurred less frequently in the testing condition ($M = 1.09$) than in the no-testing condition ($M = 2.71$). Retention interval also had a main effect, $F(1, 142) = 22.71$, $p < .01$, $\eta_p^2 = .14$. Specifically, intrusions were about three times more likely to occur ($M = 2.89$) at the 25-min interval than at the 1-min interval ($M = 0.92$). Lastly, testing and retention interval did not interact, $F(1, 142) = 1.49$, $p = .22$, $\eta_p^2 = .01$, $B_{01} = 2.18$. Most important for present purposes, it is clear from Fig. 2 that a TPNL effect on intrusions persisted across the 25-min retention interval.

Recall across lists

We now examine recall performance across the four lists for participants in the testing condition (who were the only participants tested on all four lists) using a 4 (Lists 1–4) × 2 (1 min vs. 25 min) mixed ANOVA. As expected, recall performance across lists differed depending on whether participants were in the 1-min or 25-min retention interval condition, and this impression was supported by the significant interaction between list and retention interval, $F(3, 213) = 17.82$, $p < .01$, $\eta_p^2 = .20$. To further scrutinize this interaction, we conducted separate repeated measures ANOVAs for participants in the two retention interval conditions. For participants in the 1-min interval condition, recall performance remained stable across all four lists ($M_{L1} = .70$, $M_{L2} = .72$, $M_{L3} = .72$, $M_{L4} = .70$), $F(3, 114) = 0.39$, $p = .76$, $\eta_p^2 = .01$, $B_{01} = 19.34$. This finding is consistent with the idea that interspersed testing inoculates against the buildup of proactive interference. In contrast, participants in the 25-min interval condition recalled fewer items during the test for List 4 ($M = .41$) than for Lists 1–3 ($M_{L1} = .68$, $M_{L2} = .69$, $M_{L3} = .73$), $F(3, 99) = 31.99$, $p < .01$, $\eta_p^2 = .49$. This was to be expected, as the test for List 4 was delayed by 25 min.

We now examine the ARC clustering scores for participants in the testing condition. For this analysis, we collapsed the data across the two retention intervals, given that i) the procedure was identical for Lists 1–3, and ii) our previous results indicated that retention interval did not affect the clustering of List 4 items for the tested participants. A repeated measures ANOVA showed that ARC scores rose across lists, $F(3, 216) = 4.69$, $p < .01$, $\eta_p^2 = .06$, with the ARC scores rising from 0.40 in List 1 to 0.50 in List 2, 0.60 in List 3, and 0.55 in List 4. It appears that clustering reaching asymptote by List 3 (see Table 1 for means separated by intervening tasks). An important question here is whether clustering increased because participants became increasingly aware of the categorical nature of the words as they *studied* the lists or because participants were *tested* across lists. The former hypothesis suggests that semantic organization was built across lists based on continued *exposure* to related words. In contrast, the latter hypothesis suggests that exposure alone was insufficient; rather, participants built organization across lists through *retrieval*. If the exposure hypothesis is correct, then the List 4 ARC score should not differ between participants in the testing and no-testing conditions (because both groups had been exposed to the same number of related words across lists). This is clearly not the case. In fact, the List 4 ARC score for the nontested participants ($M = .34$, at the 1-min retention interval) was similar to the List 1 ARC score for the tested participants ($M = .40$, averaged across participants in the 1-min and 25-min retention interval, but note that List 1 recall actually occurred 1 min after encoding for all tested participants), $t(107) = 0.60$, $p = .55$, $d = .12$, $B_{01} = 3.98$. This finding suggests that continued exposure to related items did not facilitate semantic organization, but retrieval practice did.

Experiment 2

In Experiment 2, we replaced the no-testing condition with a restudying condition as the control. This change was implemented to examine whether the benefit of testing on new learning was due, at least in part, to the re-exposure of the same items studied prior to new learning. Although the ARC results from Experiment 1 showed that exposure to categorized words across lists did not increase recall clustering in the absence of retrieval practice, it remains possible that *re-exposure to identical words*, rather than continued exposure to related (but different) words, was responsible for the enhanced clustering (and enhanced new learning) observed for participants in the testing condition. In the testing effect literature, researchers sometimes compare a testing condition with a no-testing condition (Chan, 2010; e.g., Chan & McDermott, 2007), and at other times compare a testing condition with a restudying condition (e.g., Roediger & Karpicke, 2006). The latter comparison has the advantage of eliminating differences in time-on-task or item re-exposure between the testing and control conditions. In a similar way, including a condition in which participants restudied the words from Lists 1–3 allowed us to examine whether the TPNL effects observed earlier were driven by re-exposure or retrieval (see also Szpunar et al., 2008; Experiment 3). Specifically, testing might have potentiated new learning of List 4 because recalling, and therefore re-encoding, the studied words enhanced one's semantic organization of the materials. This enhanced semantic organization, or better recognition of the categorized structure of the lists, might potentiate encoding of new words because it facilitated relational processing of the words in a list in the testing condition (McDaniel & Einstein, 1989; McDaniel, Einstein, & Waddill, 1990). If re-exposure to the categorized words through restudying (or retrieval practice) potentiates new learning of List 4 through enhanced semantic organization, then testing should not potentiate new learning relative to restudying, as manifested by both recall performance and clustering scores.

Method

Participants, design, materials, and procedure

Participants were 104 undergraduate students from Iowa State University. Of these, one was eliminated from analysis due to an experimenter error, one was eliminated because English was not his/her primary language, and two were eliminated because they failed to follow instructions. Therefore, 100 participants were included in the final analyses, with 33 in the *restudying*, *1*-min *interval* condition, 16 in the *testing*, *1*-min *interval* condition, 34 in the *restudying*, *25*-min *interval* condition, and 17 in the *testing*, *25*-min *interval* condition. There were fewer participants in the testing conditions than in the restudying conditions because the former were direct replications of the same conditions from Experiment 1. As will be clear from the Results to follow, the data from the present testing condition closely mirrored those from Experiment 1.

Experiment 2 used a 2 (Intervening task: testing vs. restudying) × 2 (Retention interval: 1 min vs. 25 min) between-subjects design. The materials and procedure were identical to those in Experiment 1 with the following exceptions. During the presentation of Lists 1–3, participants in the restudying condition first studied each word in a list twice (i.e., identical to Experiment 1). Next, they completed math problems

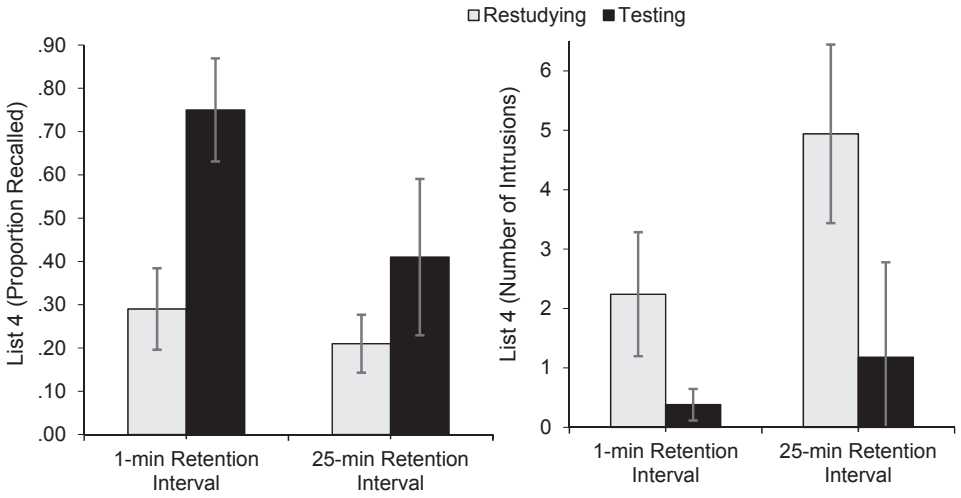


Fig. 3. Correct List 4 recall and intrusions as a function of intervening task and retention interval in Experiment 2. Left panel shows proportion of correct recall; right panel shows number of intrusions during List 4 recall. Error bars indicate descriptive 95% confidence intervals.

for 60 s, and then they restudied the words in the same list twice again, with a fresh random order for each presentation of a given list. Therefore, for Lists 1–3, participants in the restudying condition encoded each item four times, whereas participants in the testing condition encoded each item twice.[2] Most importantly, however, List 4 was presented in the same manner for all participants, such that each word was studied only twice before they were tested.

Results and discussion

List 4 recall

Correct recall. A 2 (testing vs. restudying) × 2 (1 min vs. 25 min) between-subjects ANOVA was conducted to examine the effects of testing and retention interval on the proportion of correct recall in List 4 (see left side of Fig. 3). A main effect of intervening task was found, such that participants recalled more List 4 items when they were tested on Lists 1–3 (M = .57) than when they restudied Lists 1–3 (M = .25), F(1, 96) = 37.11, p < .01, η_p^2 = .28. A main effect of retention interval was also found, F(1, 96) = 15.16, p < .01, η_p^2 = .14, which indicated that participants recalled more words after a 1-min retention interval (M = .52) than after a 25-min retention interval (M = .31). Unlike Experiment 1, however, the interaction between intervening task and retention interval was significant, F(1, 96) = 5.22, p = .03, η_p^2 = .05. At the 1-min retention interval, participants in the testing condition recalled far more words from List 4 (M = .75) than participants in the restudying condition (M = .29), t(47) = 5.88, p < .01, d = 1.79. At the 25-min retention interval, participants in the testing condition still recalled more words from List 4 (M = .41) than participants in the restudying condition, (M = .21), t(49) = 2.71, p < .01, d = .81, but the benefit of testing on new learning here was weaker than that observed at the 1-min retention interval.

Despite this finding, we believe that it might be premature to conclude that the TPNL effect had weakened across the retention interval for two reasons. First, the data of Experiment 1 indicated that the TPNL effect, relative to no-testing, persisted over the delay. It is difficult to envision why the effect would decline over time when the comparison condition was restudying instead of no-testing, given that restudying information does not typically slow forgetting (Carpenter, Pashler, Wixted, & Vul, 2008). Second, and most importantly, participants in the restudying condition exhibited very low recall performance at the 1-min retention interval. This created a situation whereby considerably less forgetting was possible in the restudying condition than in the testing condition. In other words, the significant interaction between interpolated task and retention interval might have been an artifact of the poor initial recall performance in the restudying condition.

We also note here that participants in the restudying condition recalled considerably fewer List 4 items in the 1-min retention interval condition (M = .29) than participants in the no-testing condition in Experiment 1 (M = .43), t(67) = 2.06, p = .02, d = .50. Although this restudy deficit may seem odd at first glance, it is not unusual. In fact, prior research on TPNL has reported similar patterns (e.g., Szpunar et al., 2008). We attribute this restudy deficit in new learning (relative to no-testing) to the continuous buildup of proactive interference during the encoding of Lists 1–3. Specifically, in the restudying condition, participants encoded Lists 1–3 twice as often as participants in the no-testing condition. Therefore, by the time List 4 was presented for encoding, participants in this condition had already encoded a total of six lists (although only three unique lists), whereas participants in the no-testing condition had only encoded a total of three lists. We believe that repeated studies of the first three lists might have impaired recall of List 4 relative to no-testing due to an increase in response competition.

Clustering in recall. Clustering in List 4 recall (ARC score) was examined using a 2 (testing vs. restudying) × 2 (1 min vs. 25 min) ANOVA (see the rightmost column of Table 1). The main effect of intervening task was significant, F(1, 96) = 7.10, p = .01, η_p^2 = .07, such that testing led to greater clustering (M = .53) than restudying (M = .20). Consistent with Experiment 1, retention interval had little impact on clustering, F(1, 96) = 0.13, p = .72, η_p^2 = .001, B_{01} = 4.61, with participants producing similar levels of clustering at the 1-min

[2] Because we wanted the restudy opportunity to mirror that of the original study opportunity, we presented the study list twice during the restudy trial. This procedure, however, also increased the time-on-task for participants in the restudying condition relative to the testing condition. Specifically, for Lists 1–3, participants in the restudying condition encoded the list words for 2 min, which was followed by 1 min of math, and then they restudied the list words for another 2 min. In contrast, participants in the testing condition studied the list words for 2 min, then they did 1 min of math, and then spent 1 min recalling words from that list. To ensure that our results could not be attributed to this procedural difference, we also collected data for the restudying condition under which the restudy trial presented each list word only once. Participants in this restudying condition thus spent the same amount of time on task as their tested counterparts. The data (N = 22 for the 1-min condition and N = 21 for the 25-min condition) were highly similar to those reported in the present paper (i.e., the restudy twice participants). Specifically, proportion of List 4 recall was .34 in the 1-min condition and .15 in the 25-min condition, and number of intrusions was 1.41 in the 1-min condition and 5.00 in the 25-min condition. Most importantly, consistent with the results reported in the main text, interpolated testing enhanced new learning relative to restudying in this new sample in both the 1-min condition, $t_{correct}$(36) = 4.86, $p_{correct}$ < .01, $t_{intrusion}$(36) = 2.02, $p_{intrusion}$ = .03, and the 25-min condition, $t_{correct}$(36) = 3.31, $p_{correct}$ < .01, $t_{intrusion}$(36) = 3.21, $p_{intrusion}$ < .01.

(M = .39) and 25-min retention intervals (M = .34). In addition, the interaction between intervening task and retention interval was not significant, $F(1, 96) = 0.09$, $p = .77$, $\eta_p^2 = .001$, $B_{01} = 3.23$. Once again, these data suggest that the benefits of testing (relative to restudying) persisted across the 25-min delay.

Intrusions. A 2 × 2 ANOVA was conducted to examine the effects of our independent variables on intrusions (see right side of Fig. 3). Here, a main effect was observed for intervening task, $F(1, 96) = 16.02$, $p < .01$, $\eta_p^2 = .14$, such that participants who were tested on Lists 1–3 produced fewer intrusions during List 4 recall (M = .78) than those who restudied Lists 1–3 (M = 3.59). In addition, a significant main effect was found for retention interval, $F(1, 96) = 6.19$, $p = .02$, $\eta_p^2 = .06$, with participants producing more intrusions at the 25-min retention interval (M = 3.06) than at the 1-min retention interval (M = 1.31). Similar to the results of Experiment 1, the interaction between testing and retention interval was not significant, $F(1, 96) = 1.82$, $p = .18$, $\eta_p^2 = .02$, $B_{01} = 1.42$. Once again, as can be seen clearly in Fig. 3, the intrusion data showed that the benefits of retrieval on new learning remained through the delay.

Recall across lists

Recall performance across lists for participants in the testing condition was analyzed using a 4 (List 1–4) × 2 (1 min vs. 25 min) mixed ANOVA. Similar to the results in Experiment 1, list and retention interval interacted, $F(3, 93) = 8.59$, $p < .01$, $\eta_p^2 = .22$. Whereas participants in the 1-min condition performed similarly across all lists, ($M_{L1} = .72$, $M_{L2} = .76$, $M_{L3} = .73$, $M_{L4} = .75$), $F(3, 45) = 0.44$, $p = .73$, $\eta_p^2 = .03$, $B_{01} = 7.71$, those in the 25-min condition recalled, as expected, substantially fewer words from List 4 than from the remaining lists ($M_{L1} = .69$, $M_{L2} = .68$, $M_{L3} = .73$, $M_{L4} = .41$), $F(3, 48) = 11.83$, $p < .01$, $\eta_p^2 = .43$.

Clustering (ARC scores) across lists for the tested participants was analyzed using a repeated measures ANOVA (see Table 1). We again collapsed the data across retention interval for the same reasons as described in Experiment 1. There was a marginally significant effect of list on the ARC scores, $F(3, 96) = 2.63$, $p = .05$, $\eta^2 = .08$, with the ARC score for List 1 being lower than Lists 2–4 ($M_{L1} = .29$, $M_{L2} = .54$, $M_{L3} = .57$, $M_{L4} = .53$). Similar to the results of Experiment 1, these data showed that clustering peaked by List 3, with the greatest gain observed between Lists 1 and 2.

To examine whether re-exposure to categorized words was able to increase clustering in the absence of retrieval practice, we compared the List 4 ARC scores for participants in the restudying condition with the List 1 ARC scores for participants in the testing condition. The List 4 ARC score for participants in the restudying condition (M = .20 after 1-min of math) did not differ from the List 1 ARC scores for participants in the testing condition (M = .29), $t(64) = 0.78$, $p = .44$, $d = .12$, $B_{01} = 3.05$. This finding is consistent with that from Experiment 1, and it suggests that retrieval practice, rather than repeated exposures to related words, increased semantic organization during subsequent retrieval.

Experiment 3

The purpose of Experiment 3 was to determine the effects of delaying new learning (rather than delaying the test for new learning) on the TPNL effect. Specifically, for all participants in this experiment, a 25-min lag occurred between the intervening task of List 3 and the encoding of List 4, during which participants completed brain teasers and played Tetris (the same tasks used during the retention interval in Experiments 1 and 2). We did not include a no-lag condition in this experiment because such a condition was identical to the 1-min retention interval condition in Experiments 1 and 2. It is important to note that, unlike interspersed testing, having participants do math problems between encoding episodes does not potentiate new learning (Szpunar et al., 2008; Weinstein et al., 2011; Wissman et al., 2011). At first glance, this result seems to suggest that testing is special, because other intervening activities (e.g., doing math problems) do not enhance new learning. However, there are two reasons to be cautious in drawing this conclusion. First, the lag between original learning and new learning is very short in these studies (about 1 min, Allen & Arbak, 1976; Arkes & Lyons, 1979; Nunes & Weinstein, 2012; Szpunar, Jing, & Schacter, 2014; Tulving & Watkins, 1974; Weinstein et al., 2014), which might be inadequate to serve as a study break. Second and more importantly, the prevailing wisdom emerging from the context change literature is that doing math problems does not trigger context change from encoding (Abel & Bauml, 2016; Klein, Shiffrin, & Criss, 2011; Sahakyan & Hendricks, 2012). To alleviate these concerns, we opted for lag activities that were very different from the encoding task and to substantially expand the duration of the lag from 1 min to 25 min, which was nearly double the duration of the entire encoding task for Lists 1–3. If prior episodic retrieval is necessary to enhance new learning, then a TPNL effect should be observed even when List 4 is encoded after a 25-min lag. In contrast, if interspersed testing enhances new learning because it serves a function similar to inserting a study break, then TPNL should not occur after the lag.

Method

Participants, design, materials, and procedure

Participants were 127 undergraduate students from Iowa State University. Six participants were eliminated due to an experimenter error, 10 because English was not their primary language, and one because the participant failed to follow instructions. Therefore, 110 participants were included in the final analyses, with 39 in the *no-testing* condition, 33 in the *restudying* condition, and 38 in the *testing* condition. The materials and procedure were identical to those used in Experiments 1 and 2. As depicted in Fig. 1, the only difference between Experiment 3 and Experiments 1 and 2 was that the 25-min delay preceded the *encoding* of List 4 rather than the test for List 4.[3]

Results and discussion

List 4 recall

Correct recall. A one-way between-subjects ANOVA showed a main effect of intervening task (testing, no-testing, restudying) on List 4 correct recall (see the left side of Fig. 4), $F(2, 107) = 12.22$, $p < .01$, $\eta^2 = .19$.[4] Specifically, participants in the testing condition recalled more List 4 items (M = .71) than those in the no-testing condition (M = .43), $t(75) = 5.11$, $p < .01$, $d = 1.16$, and the restudying condition (M = .50), $t(69) = 3.43$, $p < .01$, $d = .82$. No significant difference in List 4 recall was observed between the participants in the no-testing and restudying conditions, $t(70) = 1.12$, $p = .27$, $d = 0.27$, $B_{01} = 2.40$. These results show that interspersed testing enhanced new learning despite the 25-min lag, during which participants completed a series of tasks unrelated to episodic encoding. This finding suggests that a study break alone, even one as long as 25 min, does not potentiate new learning, at least when the dependent variable is correct recall. Instead, prior retrieval appears necessary to alter how participants encode and/or retrieve new information. We describe this idea in detail in the General Discussion.

[3] Similar to Experiment 2, participants in the restudying condition re-encoded the items in Lists 1–3 twice, which increased their time-on-task by 60 s per list relative to participants in the testing and no-testing conditions. To address this difference in methodology, we tested an additional group of participants (N = 23) who restudied each word only once. Proportion of correct recall was .53 and number of intrusions was 1.35 for this group of participants. Similar to the conclusion in the main text, interpolated testing enhanced List 4 recall relative to restudying, $t(59) = 2.76$, $p < .01$. However, unlike the results in the main text, interpolated testing reduced List 4 intrusions relative to restudying, $t(59) = 2.03$, $p = .02$, although the difference was modest.

[4] In a single independent variable ANOVA, η^2 is the same as η_p^2.

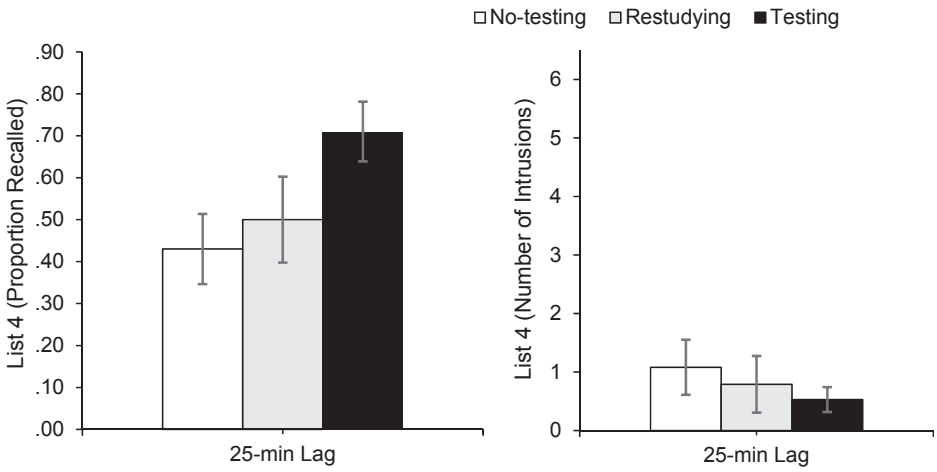


Fig. 4. Correct List 4 recall and intrusions as a function of intervening task in Experiment 3. Left panel shows proportion of correct recall; right panel shows number of intrusions during List 4 recall. Error bars indicate descriptive 95% confidence intervals.

Clustering in recall. A one-way ANOVA showed a marginal effect of intervening task on semantic clustering during List 4 recall, $F(2, 107) = 2.67$, $p = .07$, $\eta^2 = .05$ (see Table 1). Specifically, interspersed testing led to marginally greater clustering ($M = .58$) than no-testing ($M = .36$), $t(75) = 1.79$, $p = .08$, $d = 0.40$, $B_{01} = 1.07$, and significantly greater clustering than restudying ($M = .31$), $t(69) = 2.25$, $p = .03$, $d = 0.54$, $B_{01} = 2.05$.

Intrusions. Unlike the results of Experiments 1 and 2, participants exhibited few intrusions during List 4 recall regardless of the nature of the intervening task (see the right side of Fig. 4), $F(2, 107) = 2.01$, $p = .14$, $\eta^2 = .04$, $B_{01} = 2.27$. Planned comparisons showed that participants in the testing condition produced significantly fewer intrusions ($M = 0.53$) when compared to the no-testing condition ($M = 1.08$), $t(75) = 2.13$, $p = .04$, $d = 0.49$, but not when compared to the restudying condition, ($M = 0.79$), $t(69) = 1.05$, $p = .30$, $d = 0.25$, $B_{01} = 2.54$. Intrusion rates also did not differ between the no-testing and restudying conditions, $t(70) = 0.86$, $p = .39$, $d = 0.20$, $B_{01} = 2.98$. These findings contrast with those from Experiments 1 and 2, in which we observed substantially more intrusions 1 min after List 4 encoding in the no-testing ($M = 1.47$) and restudying conditions ($M = 2.24$) than in the present experiment. Therefore, although the lag had little impact on the magnitude of the TPNL effect for correct recall, it reduced the effect for intrusions.

Recall across lists

Similar to Experiments 1 and 2, recall across lists remained stable for participants in the testing condition, ($M_{L1} = .66$, $M_{L2} = .68$, $M_{L3} = .72$, $M_{L4} = .71$), $F(3, 111) = 1.29$, $p = .28$, $\eta^2 = .03$, $B_{01} = 6.41$. Once again, this result shows that the lag did not affect learning of List 4 for the tested participants.

A repeated measures ANOVA revealed that clustering increased across lists ($M_{L1} = .35$, $M_{L2} = .61$, $M_{L3} = .62$, $M_{L4} = .58$), $F(3, 111) = 4.46$, $p < .01$, $\eta^2 = .108$. Consistent with the data from Experiments 1 and 2, these ARC scores showed that clustering reach asymptotic level by List 3, with the greatest gain observed between Lists 1 and 2. In addition, the List 4 ARC score for participants in the no-testing condition ($M = .36$) and the restudying condition ($M = .31$) were comparable to the List 1 ARC score ($M = .35$) for the participants in the testing condition, $F(2, 107) = 0.09$, $p = .92$, $B_{01} = 10.96$. Once again, this finding indicates that repeated *exposure* to categorized words did not foster clustering in subsequent recall, but retrieval practice did.

Experiment 4

The results of Experiment 3 clearly showed that the benefits of retrieval on new learning remained robust despite the 25-min lag prior to new learning. However, because we did not include a short-lag condition in Experiment 3, conclusions regarding the effects of lag must be inferred on the basis of cross-experimental comparisons (i.e., against the 1-min retention interval condition in Experiments 1 and 2). Therefore, in Experiment 4, we attempted to replicate and extend the findings of Experiment 3 with the addition of a 1-min lag condition. Our objective was to compare the effects of a short- vs. a long-lag in a single experiment. To this end, we conducted an experiment in which we manipulated both lag (1 min, 25 min) and intervening task (no-testing, restudying, testing) between-subjects.

Method

Participants, design, materials, and procedure

A total of 238 participants participated in this experiment. The data from five participants were omitted from the analysis: two due to English not being their primary language, one due to an experimenter error, one due to data corruption, and one due to the participant not following instructions. The final data set therefore included 233 participants, with 36 in the *no-testing*, *1*-min *lag* condition, 42 in the *no-testing*, *25*-min *lag* condition, 36 in the *restudying*, *1*-min *lag* condition, 41 in the *restudying*, *25*-min *lag* condition, 37 in the *testing*, *1*-min *lag* condition, and 41 in the *testing*, *25*-min *lag* condition.

The procedure for Experiment 4 was identical to that of the previous experiments, except that half of the participants were in the 1-min lag conditions (similar to the short retention interval conditions in Experiments 1 and 2) and the remaining participants were in a 25-min lag conditions (similar to Experiment 3).

Results and discussion

List 4 recall

Correct recall. The data for Experiment 4 were consistent with those from Experiments 1–3, with interpolated testing producing a substantial benefit on new learning relative to both no-testing and restudying, and this benefit persisted through the 25 min lag (see Fig. 5). These impressions were supported by the results of a 3 (testing, restudying, no-testing) × 2 (1-min lag, 25-min lag) between-subjects ANOVA. Specifically, there was a main effect of interpolated task, $F(2, 227) = 41.74$, $p < .01$, $\eta_p^2 = .27$, with the tested participants recalling

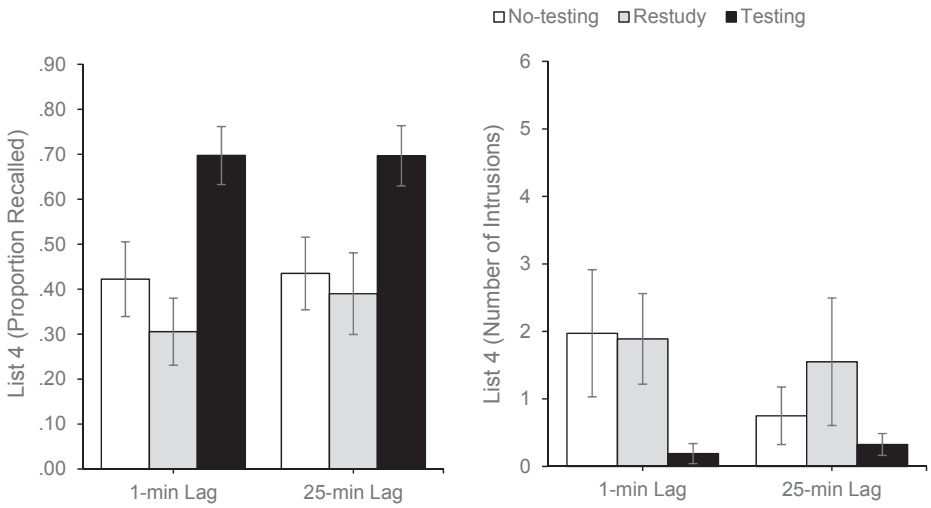


Fig. 5. Correct List 4 recall and intrusions as a function of lag and intervening task in Experiment 4. Left panel shows proportion of correct recall; right panel shows number of intrusions during List 4 recall. Error bars indicate descriptive 95% confidence intervals.

more List 4 words (M = .70) than both the nontested participants ($M_{nontested}$ = .43), $t(154) = 7.23$, $p < .01$, $d = 1.16$, and the restudied participants ($M_{restudied}$ = .36), $t(153) = 8.75$, $p < .01$, $d = 1.41$. Further, neither the main effect of lag, $F(1, 227) = 1.59$, $p = .201$, $\eta_p^2 < .01$, $B_{01} = 4.29$, nor the interaction between interpolated task and lag was significant, $F(2, 227) = 0.93$, $p = .40$, $\eta_p^2 < .01$, $B_{01} = 5.59$. Indeed, an examination of Fig. 5 shows clearly that the benefits of testing on new learning were robust in both lag conditions.

Before turning to the data on semantic clustering, we note an interesting finding that is similar to one we reported in Experiment 2. Specifically, participants in the restudying condition recalled fewer List 4 words (M = .31) than participants in the no-testing condition (M = .42) – a restudy deficit, $t(70) = 2.05$, $p = .04$, $d = 0.48$. Once again, we interpret this finding as representative of the fact that repeatedly studying the list items allowed more proactive interference to build up across list relative to studying each list only once, which further suppressed learning of List 4.

Clustering in recall. Clustering in List 4 recall (ARC score) was examined in a 3 (testing vs. restudying vs. no-testing) × 2 (1-min lag vs. 25-min lag) ANOVA (see Table 1), and the results mirrored those from correct recall. The main effect of intervening task was significant, $F(2, 212) = 12.99$, $p < .01$, $\eta_p^2 = .07$. Specifically, participants who were tested on Lists 1–3 were far more likely to cluster their recall during List 4 (M = 0.65) than participants who were not tested (M = 0.21), $t(150) = 5.24$, $p < .01$, $d = 0.85$, and participants who restudied those lists (M = 0.22), $t(142) = 4.30$, $p < .01$, $d = 0.72$. Moreover, neither lag, $F(1, 212) = 0.39$, $p = .54$, $\eta_p^2 < .01$, $B_{01} = 5.38$, nor its interaction with intervening task was significant, $F(2, 212) = 0.44$, $p = .65$, $\eta_p^2 = 0.01$, $B_{01} = 11.00$.

Intrusions. A 3 × 2 ANOVA showed a main effect of intervening task, $F(2, 227) = 10.60$, $p < .01$, $\eta_p^2 = .09$, a main effect of lag, $F(1, 227) = 3.98$, $p = .05$, $\eta_p^2 = .02$, and an interaction that was marginal, $F(2, 227) = 2.48$, $p = .09$, $\eta_p^2 = .02$. The main effect of intervening task showed that testing reduced the number of intrusions (M = 0.27) during List 4 recall relative to restudying (M = 1.66), $t(153) = 4.62$, $p < .01$, $d = 0.74$, and no-testing (M = 1.34), $t(154) = 3.90$, $p < .01$, $d = 0.62$. To further examine the effects of lag on intrusions, we conducted separate t-test for each intervening task condition. As can be seen in the right panel of Fig. 5, performing retrieval practice on Lists 1–3 nearly eliminated intrusions during List 4 recall, regardless of whether a 1-min (M = 0.19) or 25-min lag (M = 0.34) preceded the encoding of List 4, $t(76) = 1.35$, $p = .18$, $d = 0.31$, $B_{01} = 1.94$. However, increasing the lag from 1 min to 25 min reduced intrusions for participants in the no-testing condition ($M_{1\text{-}min} = 1.97$, $M_{25\text{-}min} = 0.74$), $t(76) = 2.53$, $p = .01$, $d = 0.57$, a conclusion consistent with the one from Experiment 3. In contrast, although increasing the lag also reduced intrusions for participants in the restudying condition ($M_{1\text{-}min} = 1.89$, $M_{25\text{-}min} = 1.44$), $t(75) = 0.76$, $p = .45$, $d = 0.17$. $B_{01} = 3.30$, the effect was not significant. Notably, this latter conclusion differed from that based on Experiment 3, in which participants in the restudy condition showed fewer intrusions ($M_{25\text{-}min} = 0.79$) than their counterparts in Experiment 2 ($M_{1\text{-}min} = 2.24$). We suspect that this discrepancy might simply be the result of sampling differences. To obtain a more representative result, we examined the effects of lag on intrusion for the restudy participants by combining the data from Experiments 2 (1-min lag), 3 (25-min lag), and 4 (1-min lag and 25-min lag). The outcome of this analysis revealed a significant, but modest, effect of lag on intrusions, $t(141) = 2.21$, $p = .03$, $d = 0.37$, with participants producing more intrusions during List 4 recall following a 1-min lag (M = 2.06) than following a 25-min lag (M = 1.15).

Recall across lists

Recall performance across lists for participants in the testing condition was analyzed using a 4 (List 1–4) × 2 (1-min lag vs. 25-min lag) mixed ANOVA. Similar to Experiment 3, recall probabilities remained stable across lists, $F(3, 228) = 0.23$, $p = .87$, $\eta_p^2 < .01$, $B_{01} = 51.76$. Moreover, lag had no effects on recall overall, $F(1, 76) = 0.28$, $p = .60$, $\eta_p^2 < .01$, $B_{01} = 3.35$, nor did it interact with lists, $F(3, 228) = 0.19$, $p = .90$, $\eta_p^2 < .01$, $B_{01} = 24.08$. In the 1-min lag condition, proportions recalled from List 1–4 were .68, .67, .68, and .70, respectively, and in the 25-min lag condition, they were .69, .70, .71 and .70.

We analyzed the ARC scores across lists for the tested participants using a repeated measures ANOVA. Similar to the results from Experiments 1–3, the data showed that ARC scores rose across lists, with a majority of the increase occurring between Lists 1 and 2 ($M_{L1} = 0.32$, $M_{L2} = 0.55$, $M_{L3} = 0.63$, $M_{L4} = 0.66$), $F(3, 228) = 12.15$, $p < .01$, $\eta_p^2 = .14$). Moreover, the List 4 ARC scores for participants in the no-testing condition ($M_{no\text{-}testing} = 0.22$) and the restudying condition ($M_{restudying} = 0.21$) did not differ from the List 1 ARC scores for participants in the testing condition, $ts < 1.32$, $ps > .19$, $ds < 0.19$, $B_{01}s > 2.58$. This finding, once again, suggests that neither continued exposures to the lists (i.e., by studying three inter-related lists in the no-testing condition) nor repeated exposures to the lists (i.e., by restudying each list) improved clustering during the recall of List 4.

General discussion

In four experiments, we found that interspersing retrieval practice between encoding episodes enhanced new learning relative to both a no-testing and a restudying baseline. The critical findings can be summarized as follows. First, as indicated by List 4 correct recall, testing enhanced new learning relative to no-testing and restudying, and this effect occurred regardless of the length of retention interval and lag. Second, based on the intrusion data, testing potentiated new learning relative to no-testing and restudying at both the 1-min and 25-min retention intervals. However, increasing the lag between original learning and new learning also substantially reduced intrusions for both the no-testing and restudying participants, thus reducing the advantage of testing in this regard. Third, as indicated by the ARC clustering scores, testing enhanced semantic organization during List 4 recall relative to both no-testing and restudying. Moreover, this benefit of prior retrieval on clustering scores persisted across both the 25-min retention interval and lag. We now discuss the theoretical and practical implications of these findings.

The persistence of test-potentiated new learning

As we have described in the Introduction, the true influence of retention interval (i.e., in the absence of contamination from a prior test) on the TPNL effect was previously unknown. Prior attempts at examining the persistence of the TPNL effect have typically used a repeated testing procedure, in which the delayed test for new learning (similar to the List 4 test in the present experiments) was repeated across both the shorter and longer retention intervals. Amongst these studies, some have observed nearly equivalent magnitudes of TPNL at both a 1-min and 30-min retention interval (Szpunar et al., 2008), whereas others have found the effect to diminish considerably from an immediate test to a 15-min delayed test (Wissman & Rawson, 2015). Because these studies administered the memory test for new learning over multiple occasions, it is difficult to ascribe differences in performance, or lack thereof, between the earlier and later tests to retention interval alone. Specifically, any reduction in the TPNL effect on the second test could be due to a beneficial effect of the first criterial test in the *control* conditions. Additionally, testing the new learning materials repeatedly might alter the retrieval processes that one invokes during recall. For example, Pierce et al. (2017) argued that prior testing of the original learning materials renders the new learning materials distinctive – because the new learning materials are the only items that have not yet been tested – this distinctiveness in turn facilitates post-retrieval monitoring, which allows participants to reduce intrusions in recall. To test this idea, Pierce et al. had their participants take a test for the new learning materials twice, with the second test occurring just two minutes after the first. Their logic was that any distinctiveness advantage enjoyed by the new-learning items would be removed by the first test, after which all studied items would have been tested once. Consistent with this idea, the TPNL-associated reduction in intrusions was markedly weakened in the second test. In short, testing the critical new-learning items across multiple occasions does not provide the ideal paradigm to examine the influence of retention interval on test-potentiated new learning.

In Experiments 1 and 2, we examined the persistence of the TPNL effect across a shorter (1 min) and a longer (25 min) retention interval without the contamination of repeated testing, and our results showed that retrieval potentiated new learning at both retention intervals. Notably, the benefits of testing on new learning were observed regardless of whether the baseline condition was no-testing (Experiment 1) or restudying (Experiment 2), and whether the dependent measure was accurate recall, intrusions, or semantic clustering. Across these analyses, only one (out of six) showed that the TPNL effect was significantly weaker at the 25-min retention interval than at the 1-min retention interval (i.e., List 4 correct recall in Experiment 2). We caution against over-interpreting this result because, as described earlier, this finding was likely driven by the very poor recall performance at the 1-min retention interval for the restudy participants. Moreover, when the data from the three dependent variables (i.e., correct recall, intrusions, and clustering scores) were evaluated as a whole, they showed that a robust TPNL effect can be found at both the 1-min and 25-min retention intervals. Despite these promising findings, a note of caution is in order: we manipulated retention interval at a relatively modest scale with only two time points (i.e., 1 min vs. 25 min). Consequently, our understanding of the persistence of TPNL will benefit from future investigations that include more time points and longer retention intervals.

Explaining testing-potentiated new learning

Why does testing potentiate new learning? One possibility is that a switch in context conferred by testing isolates the original learning episode (i.e., the list studied before retrieval practice) from the new learning episode (i.e., the list studied after retrieval practice), similar to the effects of taking a break from studying. Alternatively, taking a test may help participants switch to more effective strategies for future encoding and/or retrieval (Chan et al., 2017; Cho et al., 2017).

In the present study, we tested the context change account using the lag manipulation and the strategy change account with list-by-list clustering analyses. If retrieval potentiates subsequent learning because it alters task context, the lag activities should do the same, and one should not observe a significant TPNL effect in Experiment 3. Based on the correct recall and the semantic clustering data, the 25-min lag had little impact on test-potentiated new learning, as the TPNL effect remained robust despite the lag. These findings suggest that the benefits of testing on new learning go beyond simply providing a break (or a change in context) from encoding activities. A potential concern with this conclusion is that perhaps our lag manipulation failed to cause context change for participants who did not take the interpolated test. Contrary to this possibility, the *intrusion data* indicate that the 25 min lag had a beneficial effect for participants in the no-testing and restudying conditions, such that they produced fewer intrusions following the 25-min lag than following the 1-min lag. The intrusion results thus indicate that the lag was likely successful at inducing a context change because it helped participants isolate Lists 1–3 from List 4. A potential argument here is that perhaps retrieval induces a more powerful context change than activities that do not involve retrieval. However, we find this argument unconvincing due to its circularity (i.e., retrieval enhances new learning more than a study break because retrieval changes context more than a study break).

We believe that the present results, and the ARC data in particular, are consistent with the idea that prior retrieval enhances new learning because it causes participants to use superior encoding/retrieval strategies. This idea is not entirely new, as researchers have recently proposed that testing may cause learners to shift to more “efficient” or more “elaborative” encoding strategies (Cho et al., 2017; Gordon & Thomas, 2017). However, it is not yet clear what strategies are considered more efficient. With the present materials, we interpreted our results as follows: During retrieval practice (but not during restudying or no-testing), participants became sensitive to the categorical structure of the lists when they used a recalled item to cue the retrieval of other studied items (Carpenter, 2011; Chan, McDermott, & Roediger, 2006; Pyc & Rawson, 2010). For example, participants might recall the word “thunder,” which might serve as a retrieval cue for “wind.” We believe this associative cuing among retrieval candidates can happen spontaneously, which in turn *alters participants’ encoding strategy* for the upcoming study lists in two important ways. First, prior retrieval might bias participants’ encoding strategy toward detecting related words *within a study list*, which should enhance relational processing among these words (Hunt & McDaniel, 1993) and strengthen their retention. Second, this bias toward processing relational elements of the words

might increase the likelihood that related words from prior lists would be spontaneously retrieved (Hintzman, 2004), which in turn facilitates the integration of these items *across lists* (Wahlheim, 2015). Together, these mechanisms might be responsible for the test-potentiated new learning effect, at least as it pertains to the present materials. In the current paper, we refer to this explanation as a strategy change account, while acknowledging that this account incorporates ideas that are not explicitly based on changes in encoding strategy, such as recursive reminding and study-phase retrieval (Hintzman, 2009; Jacoby, Wahlheim, & Kelley, 2015; Wahlheim, 2015). To be clear, we believe that testing can induce an encoding strategy change that may cascade to other processes that are beneficial for new learning.

If enhanced relational encoding of the categorized words contributes to the TPNL effect, one should expect that the magnitude of this effect would be amenable to manipulations of presentation order of the words. Specifically, in the present experiments, we always presented words for encoding in a random order, which obscured the categorical structure of the list. If testing potentiates new learning because it facilitates relational encoding of the words, then its benefits should be reduced when words belonging to the same category are presented in blocks (e.g., consecutively). Presenting related words in blocks should encourage relational processing, thereby minimizing the difference in processing orientation between the tested and nontested participants. This prediction is borne out in a study by Nunes and Weinstein (2012), in which participants studied words from the Deese-Roediger-McDermott (DRM) associative lists (1995). In one experiment, the words from each DRM list (e.g., hill, valley, summit) were spread across five study lists. Participants either received retrieval practice after each of the first four lists or not, and then all participants were tested on List 5. Similar to the present experiments, interspersed testing promoted learning of the words in List 5. Critically, in their Experiment 2, Nunes and Weinstein presented all words of a given DRM list together in List 1–4 (instead of spreading the words across lists), such that each study list corresponded to a single DRM association (e.g., all the words in List 1 were related to *mountain*, all the words in List 2 were related to *soft*). Consistent with the strategy change account proposed here, the TPNL effect was absent with this blocked presentation method, presumably because 1) the blocked presentation no longer allows relational processing of items across lists for the tested participants or 2) the blocked presentation naturally invited relational processing within list for the nontested participants. However, an alternative explanation is also possible. Specifically, the TPNL effect might not have occurred when each study list consisted of a different set of semantic associates because presenting related words in a blocked fashion might have prevented the buildup of proactive interference in the control, non-tested condition. This logic is based on the finding that switching semantic categories during encoding can release learners from proactive interference (Wickens, 1970). But perhaps more importantly, how does changing semantic categories release learners from proactive interference? One possibility is that changing semantic categories evokes a context change (Bauml & Kliegl, 2013). But as we have discussed extensively above, we do not believe that a context change account serves as the best explanation for testing-potentiated new learning. Consequently, we argue here that the strategy change account is a more viable (and testable) explanation for both the present findings and prior findings (Nunes & Weinstein, 2012).

One may question whether this strategy change account can explain the TPNL effect in other situations, such as when participants study unrelated word lists (Aslan & Bauml, 2015; Pastotter & Bauml, 2014; Pastotter, Weber, & Bauml, 2013; Pastotter et al., 2011) or more complex materials like video lectures or text passages (Chan & LaPaglia, 2011; Gordon & Thomas, 2014; Szpunar et al., 2013; Szpunar et al., 2014; Wissman & Rawson, 2015; Wissman et al., 2011). Because unrelated words do not normally lend themselves to relational processing, they may reduce any categorical processing advantage induced by prior testing. Consequently, one may argue that the strategy change account would not predict a TPNL effect with unrelated word lists – unless one expands the idea of a strategy change to relational processing to include ad-hoc relations. For example, when performing retrieval practice of unrelated words, participants may notice or generate ad-hoc associations among these words, and they can then apply this relational encoding strategy when studying subsequent lists. The effort required to produce these ad-hoc relations would likely be greater than that needed to process the pre-existing associations for semantically related words, so the TPNL effect should be smaller with unrelated word lists than moderately related word lists. Although this prediction has not yet been tested empirically, a recent meta-analysis showed that, indeed, studies that used related words tended to show a greater TPNL effect than studies that employed unrelated words (Chan et al., in preparation).

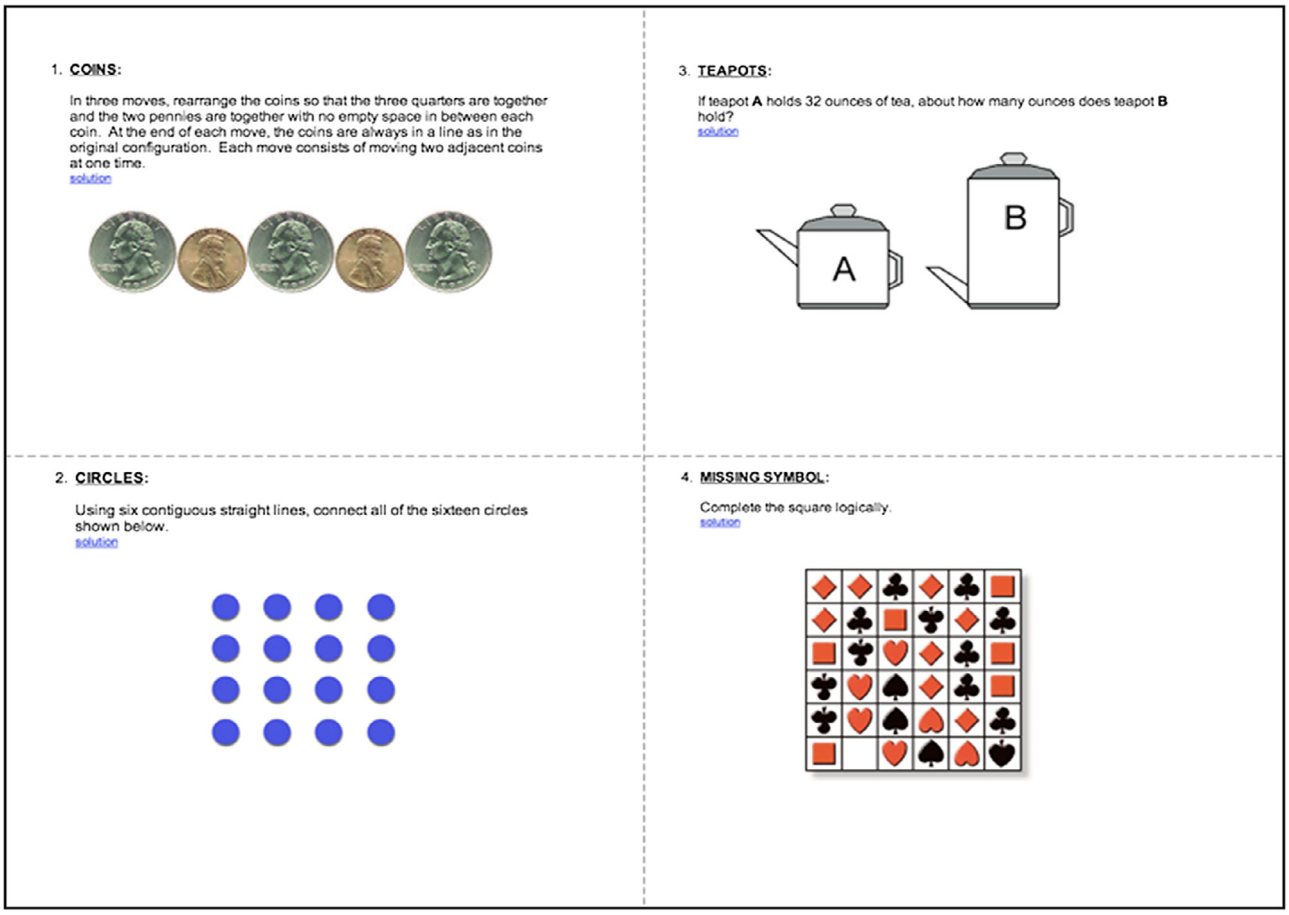
In contrast to unrelated word lists, text passages and videos are typically written/produced in a coherent manner, which should naturally invite relational processing, so any relational processing advantage induced by prior testing is likely to be modest relative to baseline (Einstein, McDaniel, Bowers, & Stevens, 1984; Einstein, McDaniel, Owen, & Cote, 1990; Masson & McDaniel, 1981). A version of the strategy change account that is not tied strictly to relational processing, however, may provide a reasonable explanation for the TPNL effect with text passages and videos. In a broader sense, the strategy change account specifies that performing retrieval practice allows participants to discover the type of learning needed to ensure satisfactory performance (or conversely, to realize the type of learning that is inadequate to produce satisfactory performance, if participants are performing poorly during retrieval practice), and participants can then adjust their subsequent encoding strategy accordingly. If we take this broader approach to strategy change, then this account can explain the TPNL effect with prose/video materials. However, we realize that the idea that "retrieval practice can improve later encoding strategies" is perhaps vaguely defined. In fact, such a broad definition of strategy change may render the account difficult to falsify. With this in mind, we believe that the strategy change account, as we currently conceive, should only be applied to explain the TPNL effect with word list type materials, for which advantageous encoding strategies can be more precisely defined (but see Jing et al., 2016 in which interspersed testing improved conceptual integration of materials across sections of a video lecture). In our opinion, application of this account to prose/video material should only be done when one clearly outlines what is considered an advantageous encoding strategy so that the hypothesis can be adequately tested.

Concluding remarks

Effective learning often requires learners to sustain their attention for prolonged periods of time – a difficult proposition (Smallwood, Fishman, & Schooler, 2007; Smallwood & Schooler, 2015). Recent research, however, has pointed to the possibility that inserting retrieval practice into an encoding task can reduce inattention and potentiate learning (Szpunar et al., 2013, for a recent review, see Szpunar, 2017). In the present experiments, we demonstrated that testing does not enhance subsequent learning simply because it provides a break from the encoding activities. Instead, performing retrieval practice changes how learners approach new, to-be-learned information, and this benefit of retrieval on new learning persists over a moderate retention interval. From a theoretical perspective, these results help shed light on the mechanisms that might be responsible for test-potentiated new learning; from a practical perspective, the present findings add to a growing literature of the multi-faceted benefits of retrieval practice on student learning.

Appendix A

A screenshot of sample brain teaser questions used in Experiments 1–4.



B. Supplementary material

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.jml.2018.05.007.

References

Abel, M., & Bauml, K. H. (2016). Retrieval practice can eliminate list method directed forgetting. *Memory & Cognition, 44*(1), 15–23. http://dx.doi.org/10.3758/s13421-015-0539-x.

Adesope, O. O., Trevisan, D. A., & Sundararajan, N. (2017). Rethinking the use of tests: A meta-analysis of practice testing. *Review of Educational Research*. http://dx.doi.org/10.3102/0034654316689306.

Allen, G., & Arbak, C. J. (1976). The priority effect in the A-B, A-C paradigm and subjects' expectations. *Journal of Verbal Learning and Verbal Behavior, 15*, 381–385.

Ariga, A., & Lleras, A. (2011). Brief and rare mental "breaks" keep you focused: Deactivation and reactivation of task goals preempt vigilance decrements. *Cognition, 118*(3), 439–443. http://dx.doi.org/10.1016/j.cognition.2010.12.007.

Arkes, H. R., & Lyons, D. J. (1979). A mediational explanation of the priority effect. *Journal of Verbal Learning and Verbal Behavior, 18*, 721–731.

Aslan, A., & Bauml, K. H. (2015). Testing enhances subsequent learning in older but not in younger elementary school children. *Developmental Science*. http://dx.doi.org/10.1111/desc.12340 [in press].

Bauml, K. H., & Kliegl, O. (2013). The critical role of retrieval processes in release from proactive interference. *Journal of Memory and Language, 68*(1), 39–53.

Bunce, D. M., Flens, E. A., & Neiles, K. Y. (2010). How long can students pay attention in class? A study of student attention decline using clickers. *Journal of Chemical Education, 87*(12), 1438–1443. http://dx.doi.org/10.1021/ed100409p.

Carpenter, S. K. (2011). Semantic information activated during retrieval contributes to later retention: Support for the mediator effectiveness hypothesis of the testing effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 37*(6), 1547–1552.

Carpenter, S. K., Pashler, H., Wixted, J. T., & Vul, E. (2008). The effects of tests on learning and forgetting. *Memory & Cognition, 36*, 438–448.

Centre for Teaching Excellence - University of Waterloo. (2012, November 7). From presenting to lecturing: adapting material for classroom delivery. Retrieved October 11, 2016, from https://uwaterloo.ca/centre-for-teaching-excellence/teaching-resources/teaching-tips/lecturing-and-presenting/delivery/adapting-material-classroom-delivery.

Chan, J. C. K. (2010). Long-term effects of testing on the recall of nontested materials. *Memory, 18*(1), 49–57. http://dx.doi.org/10.1080/09658210903405737.

Chan, J. C. K., Meissner, C. A., & Davis, S. D. (2018). Test-potentiated (new) learning: A meta-analytic review [in preparation].

Chan, J. C. K., & LaPaglia, J. A. (2011). The dark side of testing memory: repeated retrieval can enhance eyewitness suggestibility. *Journal of Experimental Psychology: Applied, 17*(4), 418–432. http://dx.doi.org/10.1037/a0025147.

Chan, J. C. K., Manley, K. D., & Lang, K. (2017). Retrieval-enhanced suggestibility: A retrospective and a new investigation. *Journal of Applied Research in Memory and Cognition, 6*, 213–229. http://dx.doi.org/10.1016/j.jarmac.2017.07.003.

Chan, J. C. K., & McDermott, K. B. (2007). The testing effect in recognition memory: a dual process account. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 33*(2), 431–437. http://dx.doi.org/10.1037/0278-7393.33.2.431.

Chan, J. C. K., McDermott, K. B., & Roediger, H. L. (2006). Retrieval-induced facilitation: initially nontested material can benefit from prior testing of related material. *Journal of Experimental Psychology: General, 135*(4), 553–571. http://dx.doi.org/10.1037/0096-3445.135.4.553.

Chan, J. C. K., Thomas, A. K., & Bulevich, J. B. (2009). Recalling a witnessed event increases eyewitness suggestibility: The reversed testing effect. *Psychological Science, 20*(1), 66–73.

Cho, K. W., Neely, J. H., Crocco, S., & Vitrano, D. (2017). Testing enhances both encoding and retrieval for both tested and untested items. *The Quarterly Journal of Experimental Psychology, 70*, 1211–1235. http://dx.doi.org/10.1080/17470218.2016.1175485.

Davis, S. D., & Chan, J. C. K. (2015). Studying on borrowed time: How does testing impair new learning? *Journal of Experimental Psychology: Learning, Memory, and Cognition,*

41(6), 1741–1754. http://dx.doi.org/10.1037/a0032377.
Davis, S. D., Chan, J. C. K., & Wilford, M. M. (2017). The dark side of interpolated testing: Frequent switching between retrieval and encoding impairs new learning. *Journal of Applied Research in Memory and Cognition*. http://dx.doi.org/10.1016/j.jarmac.2017.07.002 [in press].
Einstein, G. O., McDaniel, M. A., Bowers, C. A., & Stevens, D. T. (1984). Memory for prose: The influence of relational and proposition-specific processing. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 10*(1), 133–143.
Einstein, G. O., McDaniel, M. A., Owen, P. D., & Cote, N. C. (1990). Encoding and recall of texts: The importance of material appropriate processing. *Journal of Memory and Language, 29*, 566–581.
Finn, B., & Roediger, H. L. (2013). Interfering effects of retrieval in learning new information. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 39*(6), 1665–1681. http://dx.doi.org/10.1037/a0032377.
Gordon, L. T., & Thomas, A. K. (2014). Testing potentiates new learning in the misinformation paradigm. *Memory & Cognition, 42*(2), 186–197. http://dx.doi.org/10.3758/s13421-013-0361-2.
Gordon, L. T., & Thomas, A. K. (2017). The forward effects of testing on eyewitness memory: The tension between suggestibility and learning. *Journal of Memory and Language, 95*, 190–199. http://dx.doi.org/10.1016/j.jml.2017.04.004.
Gordon, L. T., Thomas, A. K., & Bulevich, J. B. (2015). Looking for answers in all the wrong places: How testing facilitates learning of misinformation. *Journal of Memory and Language, 83*, 140–151.
Gunter, B. (1980). Release from proactive interference with television news items: Evidence for encoding dimensions within televised news. *Journal of Experimental Psychology: Human Learning and Memory, 6*(2), 216–223. http://dx.doi.org/10.1037/0278-7393.6.2.216.
Hintzman, D. L. (2004). Judgment of frequency versus recognition confidence: Repetition and recursive reminding. *Memory & Cognition, 32*, 336–350.
Hintzman, D. L. (2009). How does repetition affect memory? Evidence from judgments of recency. *Memory & Cognition, 38*(1), 102–115. http://dx.doi.org/10.3758/MC.38.1.102.
Hunt, R. R., & McDaniel, M. A. (1993). The enigma of organization and distinctiveness. *Journal of Memory and Language, 32*, 421–445.
Jacoby, L. L., Wahlheim, C. N., & Kelley, C. M. (2015). Memory consequences of looking back to notice change: Retroactive and proactive facilitation. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 41*(5), 1282–1297. http://dx.doi.org/10.1037/xlm0000123.
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.
Jing, H. G., Szpunar, K. K., & Schacter, D. L. (2016). Interpolated testing influences focused attention and improves integration of information during a video-recorded lecture. *Journal of Experimental Psychology: Applied, 22*(3), 305–318. http://dx.doi.org/10.1037/xap0000087.
Jonker, T. R., Seli, P., & MacLeod, C. M. (2013). Putting retrieval-induced forgetting in context: An inhibition-free, context-based account. *Psychological Review, 120*(4), 852–872. http://dx.doi.org/10.1037/a0034246.
Klein, K. A., Shiffrin, R. M., & Criss, A. H. (2011). Putting context in context. In *The foundations of remembering: essays in honor of Henry L. Roediger, III* (pp. 171–190). New York: Routledge. http://dx.doi.org/10.4324/9780203837672.
Masson, M. E., & McDaniel, M. A. (1981). The role of organizational processes in long-term retention. *Journal of Experimental Psychology: Human Learning and Memory, 7*(2), 100.
McDaniel, M. A., & Einstein, G. O. (1989). Material-appropriate processing: A contextualist approach to reading and studying strategies. *Educational Psychology Review, 1*, 113–145.
McDaniel, M. A., Einstein, G. O., & Waddill, P. J. (1990). Material-appropriate processing: Implications for remediating recall deficits in students with learning disabilities. *Learning Disability Quarterly, 13*, 258–268.
McDaniel, M. A., Roediger, H. L., & McDermott, K. B. (2007). Generalizing test-enhanced learning from the laboratory to the classroom. *Psychonomic Bulletin & Review, 14*, 200–206.
Nunes, L. D., & Weinstein, Y. (2012). Testing improves true recall and protects against the build-up of proactive interference without increasing false recall. *Memory, 20*(2), 138–154. http://dx.doi.org/10.1080/09658211.2011.648198.
Olmsted, J. A. (1999). The mid-lecture break: When less is more. *Journal of Chemical Education, 76*(2–4), 525–527.
Pastotter, B., & Bauml, K. H. (2014). Retrieval practice enhances new learning: the forward effect of testing. *Frontiers in Psychology, 5*, 1–5.
Pastotter, B., Schicker, S., Niedernhuber, J., & Bauml, K. H. (2011). Retrieval during learning facilitates subsequent memory encoding. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 37*, 287–297.
Pastotter, B., Weber, J., & Bauml, K. H. (2013). Using testing to improve learning after severe traumatic brain injury. *Neuropsychology, 27*(2), 280–285. http://dx.doi.org/10.1037/a0031797.
Pierce, B. H., Gallo, D. A., & McCain, J. L. (2017). Reduced interference from memory testing: a postretrieval monitoring account. Journal of Experimental Psychology: Learning, Memory, and Cognition. http://dx.doi.org/10.1037/xlm0000377 [in press].
Pyc, M. A., & Rawson, K. A. (2010). Why testing improves memory: Mediator effectiveness hypothesis. *Science, 330*(6002), 335. http://dx.doi.org/10.1126/science.1191465 335–335.
Risko, E. F., Anderson, N., Sarwal, A., Engelhardt, M., & Kingstone, A. (2012). Everyday attention: variation in mind wandering and memory in a lecture. *Applied Cognitive Psychology, 26*(2), 234–242. http://dx.doi.org/10.1002/acp.1814.
Roediger, H. L., & Karpicke, J. D. (2006). Test-enhanced learning: Taking memory tests improves long-term retention. *Psychological Science, 17*, 249–255.
Roediger, H. L., & McDermott, K. B. (1995). Creating false memories: Remembering words not presented in lists. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 21*(4), 803–814.
Roenker, D. L., Thompson, C. P., & Brown, S. C. (1971). Comparison of measures for the estimation of clustering in free recall. *Psychological Bulletin, 76*, 45–48.
Rowland, C. A. (2014). The effect of testing versus restudy on retention: a meta-analytic review of the testing effect. *Psychological Bulletin, 140*(6), 1432–1463. http://dx.doi.org/10.1037/a0037559.
Sahakyan, L., & Hendricks, H. E. (2012). Context change and retrieval difficulty in the list-before-last paradigm. *Memory & Cognition, 40*(6), 844–860. http://dx.doi.org/10.3758/s13421-012-0198-0.
Sahakyan, L., & Kelley, C. M. (2002). A contextual change account of the directed forgetting effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 28*(6), 1064–1072. http://dx.doi.org/10.1037/0278-7393.28.6.1064.
Smallwood, J., Fishman, D. J., & Schooler, J. W. (2007). Counting the cost of an absent mind: Mind wandering as an underrecognized influence on educational performance. *Psychonomic Bulletin and Review, 14*, 230–236. http://dx.doi.org/10.3758/BF03194057.
Smallwood, J., & Schooler, J. W. (2015). The science of mind wandering: Empirically navigating the stream of consciousness. *Annual Review of Psychology, 66*(1), 487–518. http://dx.doi.org/10.1146/annurev-psych-010814-015331.
Soderstrom, N. C., & Bjork, R. A. (2014). Testing facilitates the regulation of subsequent study time. *Journal of Memory and Language, 73*, 99–115. http://dx.doi.org/10.1016/j.jml.2014.03.003.
Szpunar, K. (2017). Directing the wandering mind. *Current Directions in Psychological Science, 26*(1), 40–44. http://dx.doi.org/10.1177/0963721416670320.
Szpunar, K. K., Jing, H. G., & Schacter, D. L. (2014). Overcoming overconfidence in learning from video-recorded lectures: Implications of interpolated testing for online education. *Journal of Applied Research in Memory and Cognition, 3*(3), 161–164. http://dx.doi.org/10.1016/j.jarmac.2014.02.001.
Szpunar, K. K., Khan, N. Y., & Schacter, D. L. (2013). Interpolated memory tests reduce mind wandering and improve learning of online lectures. *Proceedings of the National Academy of Sciences of the United States of America, 110*, 6313–6317. http://dx.doi.org/10.1073/pnas.1221764110/-/DCSupplemental/pnas.201221764SI.pdf.
Szpunar, K. K., McDermott, K. B., & Roediger, H. L. (2008). Testing during study insulates against the buildup of proactive interference. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 34*(6), 1392–1399. http://dx.doi.org/10.1037/a0013082.
Tulving, E., & Watkins, M. J. (1974). On negative transfer: Effects of testing one list on the recall of another. *Journal of Verbal Learning and Verbal Behavior, 13*, 181–193.
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*(3), 289–335. http://dx.doi.org/10.1016/j.jml.2003.10.003.
Wahlheim, C. N. (2015). Testing can counteract proactive interference by integrating competing information. *Memory & Cognition, 43*(1), 27–38. http://dx.doi.org/10.3758/s13421-014-0455-5.
Weinstein, Y., Gilmore, A. W., Szpunar, K. K., & McDermott, K. B. (2014). The role of test expectancy in the build-up of proactive interference in long-term memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 40*(4), 1039–1048. http://dx.doi.org/10.1037/a0036164.
Weinstein, Y., McDermott, K. B., & Szpunar, K. K. (2011). Testing protects against proactive interference in face–name learning. *Psychonomic Bulletin & Review, 18*(3), 518–523. http://dx.doi.org/10.3758/s13423-011-0085-x.
Whiffen, J. W., & Karpicke, J. D. (2017). The role of episodic context in retrieval practice effects. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 43*, 1036–1046. http://dx.doi.org/10.1037/xlm0000379.
Wickens, D. D. (1970). Encoding categories of words: An empirical approach to meaning. *Psychological Review, 77*(1), 1–15. http://dx.doi.org/10.1037/h0028569.
Wissman, K. T., & Rawson, K. A. (2015). Grain size of recall practice for lengthy text material: fragile and mysterious effects on memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 41*(2), 439–455. http://dx.doi.org/10.1037/xlm0000047.
Wissman, K. T., Rawson, K. A., & Pyc, M. A. (2011). The interim test effect: Testing prior material can facilitate the learning of new material. *Psychonomic Bulletin & Review, 18*(6), 1140–1147. http://dx.doi.org/10.3758/s13423-011-0140-7.
Yang, C., Potts, R., & Shanks, D. R. (2017). The forward testing effect on self-regulated study time allocation and metamemory monitoring. *Journal of Experimental Psychology: Applied, 1–17*. http://dx.doi.org/10.1037/xap0000122.
Yang, C., Potts, R., & Shanks, D. R. (2018). Enhancing learning and retrieval of new information: A review of the forward testing effect. *Npj Science of Learning, 3*(1), 8.
Zaromb, F. M., & Roediger, H. L. (2010). The testing effect in free recall is associated with enhanced organizational processes. *Memory & Cognition, 38*(8), 995–1008. http://dx.doi.org/10.3758/MC.38.8.995.