Does testing enhance new learning because it insulates against proactive interference?

Dahwi Ahn · Jason C. K. Chan

Abstract
Taking a test on previously learned material can enhance new learning. One explanation for this forward testing effect is that retrieval inoculates learners from proactive interference (PI). Although this release-from-PI account has received considerable empirical support, most extant evidence is correlational rather than causal. We tested this account by manipulating the level of PI that participants experience as they studied several lists while receiving interpolated tests or not. In Experiments 1 and 2, we found that testing benefited new learning similarly regardless of PI level. These results contradict those from Nunes and Weinstein (Memory, 20(2), 138–154, 2012), who found no forward testing effect when encoding conditions minimized PI. In Experiments 3 and 4, we failed to replicate their results. Together, our data indicate that reduced PI might be a byproduct, rather than a causal factor, of the forward testing effect.

Keywords Retrieval practice · New learning · Forward testing effect · Proactive interference

Introduction
Many students and teachers consider tests as an assessment tool (Bonner, 2009; Karpicke et al., 2009). However, a large body of research has established that taking tests enhances learning (Chan, Meissner et al., 2018; Karpicke & Roediger, 2008; Rowland, 2014; for exceptions, see Chan et al., 2017; Davis & Chan, 2015; Finn & Roediger, 2013). The beneficial effects of testing can be classified in two ways. On the one hand, testing promotes learning of tested material, which is sometimes referred to as the backward testing effect. On the other hand, testing of previously studied material improves subsequent learning of new material, which is sometimes termed the forward testing effect. This latter effect is the focus of the current study.

The forward testing effect is often examined using a multi-list learning, interpolated testing paradigm. At the beginning of the study, participants are told that they might or might not take a test after studying each list. But in actuality, participants in the tested condition always take a test after each list, whereas those in the control condition either restudy the list or do a filler task such as solving math problems. All participants are tested on the final, target list (Fig. 1 shows the design of our experiments, but they also conform to the canonical design used in the literature). If the tested participants perform better on the target test than the control participants, then interpolated testing has facilitated new learning (i.e., a forward testing effect). The forward testing effect is robust and has been observed with a variety of conditions (for a review, see Chan, Manley et al., 2018).

Although there is a rich literature on the forward testing effect, most studies have been empirically oriented (cf. Chan et al., 2020; Chan, Manley et al., 2018; Kliegl & Bäuml, 2021). Recently, researchers have started to explore the mechanisms underlying the effect (for a review, see Chan, Meissner et al., 2018; Pastötter & Bäuml, 2014; Yang et al., 2018), but the amount of empirical studies far outstripped that of theoretical ones. Thus, the goal of the current study was to extend the theoretical understanding of the forward testing effect. Specifically, we provided a critical test of the release from proactive interference explanation (Szpunar et al., 2008; Yang et al., 2018), which suggests that interpolated testing facilitates subsequent learning because it insulates participants against the buildup of proactive interference (PI) from prior learning. Below, we review the relevant literature and explain the motivation of our study.
Proactive interference and forward testing effect

Proactive interference plays an important role in student learning. Suppose that some students are taking Spanish and French courses on the same day. During the Spanish class, they learned that the word *amor* means love. A few hours later, during the French class, they learned that *amour* (with a “u”) means love. Later, if students are asked to write love in French, they might recall *amor* instead of *amour*. This example shows one scenario in which prior learning (of Spanish) interferes with new learning (of French), i.e., proactive interference (PI).

PI is widely considered a major source of forgetting (for reviews, see Crowder, 1976; Kliegl & Bäuml, 2021). In one of the earliest demonstrations of the forward testing effect, Tulving and Watkins (1974) considered insulation of PI as a causal factor for the effect. In their study, participants studied two lists (the first containing A-B associations and the second containing A-C associations) and then attempted to recall words from both lists in a modified-modified free recall (MMFR) test. The key manipulation was whether participants were tested on the first, A-B list before they learned the second, A-C list (i.e., the presence or absence of interpolated testing), and their results showed that immediate testing of List 1 greatly increased recall of List 2. Tulving and Watkins interpreted this finding as showing that “the interpolation of an activity requiring explicit retrieval of stored information seems to insulate the A-B list from the A-C list in a way that removes the former as an interfering component in the learning of the latter” (p. 192).

Response competition is a mechanism through which PI impairs subsequent learning. Specifically, compared to a situation in which participants learn only a target list, prior learning increases the number of retrieval candidates for a given cue, and these items compete with each other and interfere with recall of the target item(s). A broadly accepted way to measure the level of response competition and PI is through intrusions of items from the nontarget lists (s) during recall testing (Bäuml & Kliegl, 2013; Chan, Meissner et al., 2018; Crowder, 1976; Darley & Murdock, 1971; Pierce et al., 2017; Postman & Hasher, 1972; Underwood, 1975; Yang et al., 2018; Zaromb & Roediger, 2010).

As evidence that interpolated testing promotes new learning by inoculating learners against the influence of PI, numerous studies have shown that interpolated testing can reduce intrusions from prior learning. For example, Darley and Murdock (1971) found that learners are far less likely to commit intrusions of previously tested items than previously nontested ones. In their study, participants studied ten lists of words while expecting a cumulative test. After studying each list, participants took an immediate recall test for half of the lists and counted forward by threes for the other lists. When examining the frequency of prior-list intrusions during the immediate-recall tests, Darley and Murdock found that most of the intrusions (i.e., 81%) belonged to the nontested lists, which shows that tested items might be less susceptible to
response competition than nontested items. This finding can be interpreted as evidence for either an early selection advantage (i.e., the tested items are less likely to come to mind when participants are attempting to retrieve the just-studied, nontested items) or a late correction advantage (i.e., participants can readily withhold a retrieved item if they remember that the item had been recalled). Either way, both possibilities suggest that the benefit of testing might stem from a reduction in PI.

**Release-from-PI account**

In a study that revived interest in the phenomenon of the forward testing effect, Szpunar et al. (2008) showed that after studying five inter-related word lists, participants who had been tested on Lists 1–4 committed fewer intrusions during List 5 recall than participants who had not been tested on Lists 1–4. Moreover, the tested participants recalled more List 5 items than the control participants. Based on these results, Szpunar and colleagues proposed the *release-from-PI account* – such that testing inoculated participants against the buildup of PI, and this was in turn responsible for the learning benefits of List 5 amongst the tested participants. Other studies have also reported similar findings that intermittent testing reduced intrusions and boosted new learning (Chan, Manley et al., 2018; Weinstein et al., 2011, 2014; Yang et al., 2018).

Because PI is theoretically neutral, this *release-from-PI account* leaves open the candidate mechanisms through which interpolated testing enhances new learning. PI can degrade subsequent learning via processes that operate either at retrieval or at encoding. In this paper, we focused on the *retrieval-based explanation* of PI for the forward testing effect. Szpunar et al. (2008) ascribed the forward testing effect to an enhancement in list discrimination – the idea is that testing boosts retention of the tested items, which in turn allowed participants to better discriminate list membership of the studied items and to avoid intrusions (see also Chan & McDermott, 2007; Jang & Huber, 2008; Postman et al., 1968; Postman & Keppel, 1977) – consequently, the original conceptualization of the release-from-PI account focused on retrieval processes, and we did the same here.

If interpolated testing enhances new learning because it reduces PI, then a forward testing benefit for correct recall should be accompanied by a reduction of intrusions. Consistent with this idea, in a recent meta-analysis (Chan, Meissner et al., 2018), out of 38 studies in which a forward testing effect was observed and intrusions were reported, 36 found that interpolated testing reduced intrusions, thus giving weight to the release-from-PI account. However, this reasoning is based on a correlation (indeed, across studies, there is a robust correlation between the magnitude of the forward testing effect in correct recall and intrusions in Chan et al.’s data, $r = -.60, p < .001$), and few studies have provided a direct test of the release-from-PI account. Specifically, if interpolated testing potentiates new learning by reducing the influence of PI, then the magnitude of the forward testing effect should depend on the level of PI experienced by learners – that is, the benefits of testing should be more pronounced when learners are expected to experience greater PI.

Szpunar et al. (2008) provided a direct test of the release-from-PI account in their Experiment 2. Specifically, they manipulated PI by varying the number of nontested lists that participants had to study before the target list (see Fig. S1 of the Online Supplemental Material (OSM) at https://osf.io/wghbc/ for their design), and PI was designed to build up across lists. Participants in the test-every-list condition should be less susceptible to the negative effects of PI. In contrast, those in the control conditions were tested on only one list among Lists 2–5 (e.g., one group was tested on only List 2, one group on List 3, and so on), and the level of PI was expected to be proportional to the number of nontested lists before testing. Consistent with the assumption that prior testing reduces the buildup of PI, when comparing the number of intrusions produced by participants in the control conditions to participants in the test-every-list condition, Szpunar et al. found that intrusions rose as the number of nontested lists increased, and this rise in intrusions was accompanied by an increase in the forward testing effect for correct recall.

Although Szpunar et al.’s (2008) results supported a release-from-PI account, manipulating the level of PI by the number of study lists raises an interpretation problem because this PI manipulation is confounded with test frequency. When comparing participants in the test-every-list condition to participants in the control condition, the benefits of interpolated testing was greater when participants in the control condition had studied more nontested lists. One interpretation of this finding is that interpolated testing reduces the influence of response competition during recall, which leads to a forward testing effect. However, other explanations are equally viable. For example, semantic clustering in recall, which serves as an indication of memory organization, increases with test frequency (Chan et al., 2020; Zaromb & Roediger, 2010), and this finding has typically been attributed to strategy optimization. Further, increasing the number of interpolated tests can promote attentive encoding because the frequent occurrence of testing signals to participants that they will likely be tested again in the near future – such that participants experience a rise in test expectancy. Critically, neither of these accounts requires release-from-PI as an explanatory mechanism. In our review of the literature, all studies (except the Nunes & Weinstein, 2012, study described below) that have provided support for the release-from-PI account suffer from similar
issues, so extant data do not provide unequivocal support for the release-from-PI account.

**PI manipulation with content change method**

In the present study, we aimed to isolate our PI manipulation from test frequency. To this end, we varied PI using the content change method in the traditional Brown-Peterson paradigm (Brown, 1958; Peterson & Peterson, 1959). Briefly, PI is accumulated when participants study a series of similar items (e.g., words from the same category, such as fruits), but when the content of the study item changes (e.g., to words in a different category, such as animals), learners experience a release from proactive interference because they can use the change in semantic category to constrain their memory search during retrieval (e.g., by attempting to retrieve only animals, Wickens et al., 1963, 1981).

The content change paradigm is ideally suited to investigate the release-from-PI account for two reasons. First, this method allowed us to manipulate the level of PI that participants experience without altering the amount of information that they had to study. More importantly, the content change method provides a way to vary PI in categorized lists without changing the words that participants study across conditions. That is, all participants studied the same words, but PI was manipulated via presentation order. Second, because our theoretical objective is to investigate whether release-from-PI, as a retrieval-based phenomenon, is responsible for the occurrence of the forward testing effect, it is particularly promising that content change has been identified as a procedure that specifically affects the retrieval component of PI effects (Engle, 1975; Gardiner et al., 1972; Kliegl & Bäuml, 2021; Wixted & Rohrer, 1993).

Using the content change method, Nunes and Weinstein (2012) reported, to our knowledge, the only causal evidence for the occurrence of the forward testing effect, it is particularly promising that content change has been identified as a procedure that specifically affects the retrieval component of PI effects (Engle, 1975; Gardiner et al., 1972; Kliegl & Bäuml, 2021; Wixted & Rohrer, 1993).

In our Experiments 1 and 2, we manipulated PI and intervening task in a factorial, between-subjects design (see Fig. 1). Specifically, PI was manipulated by presenting words from participants in the tested and control conditions recalled virtually identical proportions of the target list words – there was no forward testing effect. Furthermore, participants in both conditions produced very few intrusions. According to Nunes and Weinstein, when each list contained a single DRM theme, the switch in semantic content across lists released participants from the negative influence of PI. If interpolated testing potentiates new learning because it releases learners from PI, and if content change minimizes PI buildup in the control condition, then the benefits of interpolated testing become irrelevant and the forward testing effect should be eliminated. This was what happened in Nunes and Weinstein’s Experiment 2, thus providing strong support for the release-from-PI account.

To our knowledge, Nunes and Weinstein (2012) is the only study that had examined the release-from-PI account using a content change method. However, several characteristics of their experiments made it difficult to draw strong conclusions. First, because Nunes and Weinstein did not manipulate presentation order in the same experiment, a cross-experimental comparison between Experiments 1 and 2 is required to determine whether a reduction in PI weakened the forward testing effect. Second, the main conclusion of Nunes and Weinstein’s study hinges on a null effect (i.e., no forward testing effect) in their Experiment 2, but that experiment had only 35 participants and the critical comparison was between-subjects. Lastly, the materials differed across their Experiments 1 and 2 (e.g., each list comprised 18 words in Experiment 1 but only 12 words in Experiment 2), and Nunes and Weinstein did not fully counterbalance their words across study lists, such that the words that appeared in the target list never appeared in other lists. The lack of counterbalancing limited the generalizability of the results – it is possible that instead of the Low-PI manipulation eliminating the forward testing effect in Experiment 2, it was the specific words that appeared in List 5 that were, for some reason, not amenable to demonstrating a forward testing effect. Although this possibility might seem remote, it deserves attention because Nunes and Weinstein’s target list featured completely different words across their Experiments 1 and 2, which compounded the difficulty in drawing conclusions based on this cross-experimental comparison.

**The current study**

In our Experiments 1 and 2, we manipulated PI and intervening task in a factorial, between-subjects design (see Fig. 1). Specifically, PI was manipulated by presenting words from

---

1. There were 19 participants in the tested condition and 16 in the control condition. Their Experiment 1 had 62 participants, with 31 each in the tested and control conditions, respectively.
the same category across lists (High-PI) or within lists (Low-PI), and intervening task was manipulated by giving participants an interpolated test or interpolated restudy phase after each nontarget list. All participants took a test on the final, target list. In the High-PI condition, participants studied mixed lists comprising words from several categories, with different words from those same categories appearing in every list, thereby providing an opportunity for PI to accumulate. In the Low-PI condition, participants studied blocked lists containing words from a unique category for Lists 1–4, and the shift in category across lists was designed to release the buildup of PI. If the forward testing effect emerges because testing insulates the buildup of proactive interference, we should observe a smaller forward testing effect in the Low-PI than in the High-PI condition. In fact, based on the data from Nunes and Weinstein’s Experiment 2 (2012), one might expect that the forward testing effect would be eliminated in the Low-PI condition, because the content change manipulation should negate the PI-reducing advantage of interpolated testing. To preview, contrary to this prediction, we found a robust forward testing effect in both the High-PI and Low-PI conditions, and the magnitude of the effect was unaffected by the PI manipulation. Thus, we aimed to directly replicate Nunes and Weinstein’s Experiment 2 in our Experiments 3 and 4.

**Experiment 1**

**Method**

**Design and participants**

Participants were 119 undergraduate students from Iowa State University who completed the experiment for course credits. There were 29 participants in the High-PI, Test condition, 30 in the High-PI, Restudy condition, 32 in the Low-PI, Test condition, and 30 in the Low-PI, Restudy condition. Data from 14 participants were not analyzed because three (i.e., one in the High-PI, Test condition; two in the Low-PI, Restudy condition) did not follow instructions, and a programming error made the data for 11 participants unusable (i.e., six in the Low-PI, Test condition; five in the Low-PI, Restudy condition). The final sample included data from 107 participants, with 28 in the High-PI, Test condition, 30 in the High-PI, Restudy condition, 26 in the Low-PI, Test condition, and 23 in the Low-PI, Restudy condition. We conducted a sensitivity power analysis using G*Power (Faul et al., 2007) to compute the effect size that can be detected based on our collected sample size. We used the Low-PI condition (combined n = 49) for this analysis because this condition had fewer participants than the High-PI condition. The minimum effect size that can be detected with .80 power was d = 0.82, which was comparable to the meta-analytic effect size (d = 0.75, Chan, Meissner et al., 2018). As will become evident, inadequate statistical power was never an issue in our experiments, as our observed effect sizes always exceeded this minimum detectable effect size (at .80 power) regardless of whether participants were in the High- or Low-PI condition.

**Materials and procedure**

For Experiments 1 and 2, we opted to employ materials that are similar to the majority of studies that have demonstrated a forward testing effect. To that end, we selected words from four categories (i.e., fruits, animals, body parts, and sports) that each had at least 20 exemplars in the Van Overschelde, Rawson, and Dunlosky norms (2004). The average typicality ratings did not differ across categories \((M_{fruits} = .18, M_{animals} = .18, M_{bodyparts} = .18, M_{sports} = .18)\), \(F(3, 76) = 0.01, p = .999, R_{B} = 14.45\).

Participants studied five 16-word lists. In the High-PI condition, each list contained four exemplars from each of the four categories. The average typicality ratings did not differ across lists (range = .17 – .18), \(F(4, 75) = 0.01, p = .999, R_{B} = 19.44\). All words within a list were presented in a pseudo-random order, with items from the same category never presented consecutively. In the Low-PI condition, each of the first four lists contained words from a single category presented in a random order (e.g., List 1 might contain only fruits, List 2 might contain only animals). Most importantly, List 5 was identical for participants in both the Low-PI and High-PI conditions, and it contained four new words from each of the four studied categories. Therefore, PI was manipulated by the structure of Lists 1–4. See the top of Appendix A for the full set of materials.

Participants in groups of up to eight completed the experiment at computers separated by dividers. Participants were informed that they would study several lists of 16 words, with each word presented twice. They were also told that after studying each list, they would solve some math problems, and then they would either be asked to recall the just-presented list or to restudy the list. Critically, participants were informed that the computer randomly determined whether a test will be given following each list, so that they might be “asked to recall words after every list, after only one or two lists, or never at all.” In actuality, participants either received a test for every list (test condition) or they were only tested on List 5 (restudy condition). To ensure that participants would study each list intentionally, they were further told that there was a cumulative recall test for all lists regardless of how many tests they received.

At the start of each list, a 2-s prompt informed participants of the list number (e.g., “This is Word List 1”), which was followed by a 1-s fixation cross. Next, the words were...
is likely that this participant did not follow the instructions. For example, if a participant recalled eight fur-

Presentations were debriefed. Participants completed a short survey on demographics and course credit-granting purposes and were not scored. Finally, participants were given 60 s to recall the words from List 5. A forward testing effect is found if participants in the test condition recall more List 5 words than participants in the restudy condition. Participants then completed a Reading Span task (Conway et al., 2005) as a filler activity for 20 min and then took the cumulative recall test on Lists 1–5. The RSPAN and cumulative recall tasks were included to ensure that the experiment lasted approximately 40 min for course credit-granting purposes and were not scored. Finally, participants completed a short survey on demographics and were debriefed.

Results

For completeness’ sake, proportions of correct recall for the nontarget lists are reported in Table 1. But the focus of our research goal was target list recall. For all experiments, we first report frequency of intrusions during target list recall, which provides a measure of PI. We defined intrusions as recall of words from any nontarget lists. Extralist intrusions were not considered because they do not clearly represent proactive interference from prior list learning. If the PI manipulation was successful, there should be fewer intrusions in the Low-PI condition than in the High-PI condition.

Across all experiments, participants whose intrusions were greater than 2 SD from the mean were considered outliers. Given that we explicitly told participants to recall words from only the most recent list, those who committed a very high number of intrusions might not have followed the instructions. For example, if a participant recalled eight furniture words right after studying the animal list, we think it is likely that this participant did not follow the instructions.

Table 1 Nontarget, interpolated test performance across Experiments 1–4

|                | List 1 | List 2 | List 3 | List 4 |
|----------------|-------|-------|-------|-------|
| Experiment 1   | .63(±.16) | .63(±.18) | .65(±.17) | .64(±.17) |
| Experiment 2   | .63(±.18) | .68(±.18) | .68(±.18) |       |
| Experiment 3   | .57(±.17) | .56(±.16) | .58(±.14) | .59(±.20) |
| Experiment 4   | .63(±.17) | .62(±.20) | .63(±.22) | .65(±.22) |

Note. Standard deviations are in parentheses.

To be comprehensive, we conducted all of our analyses with and without outliers. We report results without outliers as the default, but also highlight when inclusion of the outliers led to different conclusions (to preview, only one analysis in Experiment 3 produced a different result). In addition, the full set of data are on the Open Science Framework at https://osf.io/wghbc/ and Figs. S2 and S3 of the OSM present results including outliers. None of the experiments were pre-registered.

After reporting intrusion data, we report target list correct recall. We examined whether PI level affected the magnitude of the forward testing effect. If reduced PI plays a causal role, we should observe a smaller forward testing effect in the Low-PI condition than in the High-PI condition. We also examined the correlation between intrusions and correct recall in the target list. We expect that the two variables should be negatively related such that participants with fewer intrusions should show greater correct recall — note, however, this correlation does not imply causation.

For statistical inferences, we used two-tailed tests with α = .05. We also report Bayes factors. When a result is significant, we report B10, which indicates support for the alternative hypothesis; when a result is not significant, we report B01, which indicates support for the null hypothesis. We took this approach so that a larger Bayes factor always indicates more support for the effect (null or otherwise) under consideration. All Bayesian analyses were performed using the default priors (JASP team, 2020).

Intrusions during List 5 recall

Across the entire sample, mean frequency of intrusions was 1.03 and standard deviation was 1.84. Accordingly, data from five participants who committed more than 4.71 intrusions (i.e., 1.03 + 2*1.84 = 4.71) were considered outliers and excluded from analyses. The left panel of Fig. 2 shows the intrusion data. We conducted a 2 (intervening task: test vs. restudy) × 2 (PI level: high vs. low) ANOVA with frequency of intrusions during List 5 recall as the dependent variable. There was a significant main effect of intervening task, F(1,98) = 34.84, p < .001, d = 1.14, B10 = 108753.94, such that the restudied participants (M = 1.31) committed more intrusions than the tested participants (M = 0.19). It is not surprising that interpolated retrieval reduces intrusions, as the effect has been widely demonstrated (Chan, Manley et al., 2018; Szpunar et al., 2008; Weinstein et al., 2011). More important for present purposes, the main effect of PI level was significant, F(1,98) = 12.27, p = .002, d = 0.59, B10 = 9.36, with participants in the High-PI condition (M = 1.02) committing nearly triple the intrusions relative to participants in the Low-PI condition (M = 0.38), which shows that our manipulation of PI was successful.
The interaction between intervening task and PI level was also significant, \( F(1, 98) = 5.61, p = .020, \eta_p^2 = .04, \) \( B_{10} = 2.73. \) Specifically, the tested participants produced minimal intrusions regardless of PI level (\( M_{\text{High-PI}} = 0.29 \) vs. \( M_{\text{Low-PI}} = 0.08)), \( t(52) = 1.50, p = .139, d = 0.41, B_{01} = 1.45, \) but the restudied participants produced significantly fewer intrusions if they were in the Low-PI condition than in the High-PI condition (\( M_{\text{High-PI}} = 1.81 \) vs. \( M_{\text{Low-PI}} = 0.73)), \( t(46) = 3.02, p = .004, d = 0.87, B_{01} = 9.70. \)

### List 5 correct recall

The right panel of Fig. 2 shows the List 5 correct recall data. The same \( 2 \times 2 \) ANOVA showed a main effect for intervening task, \( F(1, 98) = 15.95, p < .001, d = 0.81, B_{10} = 239.75, \) which shows a forward testing effect overall (\( M_{\text{test}} = .51 \) vs. \( M_{\text{restudy}} = .34. \) PI manipulation did not affect correct recall (\( M_{\text{High-PI}} = .43 \) vs. \( M_{\text{Low-PI}} = .44), F(1, 98) = 0.05, p = .826, d = 0.05, B_{01} = 4.64. \) Of greater interest, an interaction effect was also not significant, \( F(1, 98) = 0.78, p = .379, \eta_p^2 = .01, B_{01} = 2.95. \) Contrary to the expectation that reducing PI would also reduce the magnitude of the forward testing effect, the test group vastly outperformed the restudy group in both the Low-PI condition (\( M_{\text{test}} = .50 \) vs. \( M_{\text{restudy}} = .36), t(46) = 2.78, p = .008, d = 0.81, B_{10} = 5.88, \) and the High-PI condition (\( M_{\text{test}} = .52 \) vs. \( M_{\text{restudy}} = .32), t(52) = 3.00, p < .001, d = 0.82, B_{10} = 9.76. \) In fact, the magnitude of the forward testing effect (i.e., the effect size) was nearly identical regardless of PI condition. The robust forward testing effect in the Low-PI condition was clearly in contradiction to Nunes and Weinstein’s Experiment 2 data. Although the manipulation of PI did not affect the forward testing effect, as predicted by the release-from-PI account, we observed a substantial negative correlation between intrusions and correct recall, \( r(100) = -.45, p < .001, B_{10} = 8154.72, \) such that individuals who recalled more correct items also produced fewer intrusions (see Fig. 3). Closer examinations revealed that this correlation was largely driven by participants in the restudy condition, \( r(46) = -.45, p < .001, B_{10} = 25.82, \) as participants in the test condition had near-floor level of intrusions, thereby restricting the range of scores on this variable and minimizing the likelihood of finding any correlations, \( r(52) = -.15, p = .284, B_{01} = 3.37. \)

### Discussion

The purpose of Experiment 1 was to examine whether the magnitude of the forward testing effect is susceptible to differences in proactive inference. As expected, participants in the High-PI condition produced more intrusions than their Low-PI counterparts. Further, the effects of PI were alleviated by interpolated testing. In sum, the intrusion data validated our PI manipulation. Despite these results, neither proportion of correct recall nor the magnitude of the forward testing effect were influenced by the PI manipulation, which suggests that release from proactive interference might be a...
corollary, rather than the cause, of the forward testing effect. An additional intriguing result was the robust negative correlation between correct recall and intrusions, even though only the latter was affected by PI. At face value, this correlation is consistent with the release-from-PI account because participants who experience less PI showed greater correct recall, so testing might enhance new learning because it reduces the buildup of PI. We consider the implications of this correlation more fully in the General discussion, but suffice it to say that our data indicate that the negative association between correct recall and intrusion does not imply causation. Rather, the coupling of these variables might be driven by a third variable, such as increased attention or enhanced memory organization due to testing (Chan et al., 2020).

Although the data in Experiment 1 show that reduction in PI might not play a causal role in the forward testing effect, it remained possible that we observed a forward testing effect in the Low-PI condition because our manipulation of PI was too weak. Specifically, participants in the Low-PI, Restudy condition might have still experienced considerably greater PI than their tested counterparts, because List 5 contained words from the categories that appeared in Lists 1–4. This possibility can be examined by comparing the intrusion data between the tested and the restudied participants in the Low-PI conditions. Even when PI was intended to be low, the tested participants (M = 0.08) committed far fewer intrusions than the restudied participants (M = 0.73), t(46) = 3.10, p = .003, d = 0.90, B_{10} = 11.58. This finding shows that our manipulation did not render the PI-reducing power of testing irrelevant, which might explain why we found a forward testing effect in the Low-PI condition.

In Experiment 2, we aimed to further reduce PI for participants in the Low-PI condition. To this end, every list, including the target list, included only words from a single category. The logic is to minimize the influence of PI on target list learning and recall by eliminating the possibility that words in the target list would remind participants of the previously studied lists.

**Experiment 2**

**Method**

**Design and participants**

PI level (high vs. low) and Intervening task (test vs. restudy) were manipulated between subjects. Participants were 120 undergraduate students from Iowa State University who participated for course credits. Data from five participants were not analyzed because three did not follow instructions and two involved a computer error. The final sample thus included data from 115 participants, with 31 in the High-PI, Restudy condition, 28 in the High-PI, Test condition, 28 in the Low-PI, Test condition, and 28 in the Low-PI, Restudy condition. For the sensitivity power analysis, we used the sample size of the Low-PI condition, which had fewer participants. The analysis showed that we could detect an effect size of $d = .76$ with .80 power.

**Materials and procedure**

The procedure of Experiment 2 was identical to Experiment 1 with the following exceptions. First, in the High-PI condition, three words from each category were spread across four lists, such that each list contained words from the same four categories. In the Low-PI condition, every list, including the target list, contained words from a single category. Note that all participants studied the same words from the four categories, with the only difference across conditions being the presentation order of the words across lists. The full set of materials is presented in Appendix A. Second, words from the categories of fruits, animals, body parts, and furniture were used, and they had the same average typicality ratings ($M_{fruits} = .18$, $M_{animals} = .18$, $M_{bodyparts} = .18$, $M_{furniture} = .18$), $F(3, 60) = 0.00$, $p = 1.00$, $B_{90} = 11.71$. Lastly, participants studied four lists of 16 words each instead of five lists. Experiment 1 was conducted after Experiment 2 chronologically. But the studies were presented in the current order for exposition purposes. The logic was to first present the study with the weaker PI manipulation (i.e., Experiment 1), and then present the study with the stronger PI manipulation (i.e., Experiment 2). The number of study lists and materials differed across the two experiments because we needed more words in Experiment 1 to create List 5 that included the category words from Lists 1–4 while maintaining list length at 16 words, and Furniture does not have enough exemplars for this purpose.

**Results**

**Intrusions during List 4 recall**

The outlier cutoff for this experiment was 4.79 intrusions ($M = 0.85$, $SD = 1.97$), and data from seven participants were excluded. The left panel of Fig. 4 shows the frequency of intrusions during List 4 recall. Because participants in the Low-PI, Test condition did not commit any intrusions at all, it was not possible to conduct a $2 \times 2$ (High-PI vs. Low-PI) ANOVA. Thus, we conducted a $t$-test between the High- and Low-PI conditions to examine the influence of PI manipulation on overall intrusion rates. As expected, participants in the High-PI condition ($M = 0.71$) produced more intrusions than those in the Low-PI condition ($M = 0.11$), $t(106) = 3.46$, $p < .001$, $d = 0.67$, $B_{10} = 36.67$. 
Indeed, the frequency of intrusions in the Low-PI condition did not differ from zero, $t(52) = 1.43, p = .159, B_{01} = 2.57$. The fact that intrusions were almost entirely removed in the Low-PI condition ($M = 0.24$ for participants in the restudy condition, with 23 out of 25 participants producing no intrusions, and the remaining two producing three intrusions each; see Fig. 4) showed that our PI manipulation in Experiment 2 had the intended effect. Lastly, overall, the restudied participants ($M = 0.09$) committed more intrusions than the tested participants ($M = 0.09$), $t(106) = 4.24, p < .001, d = 0.82, B_{10} = 425.83$.

List 4 correct recall

The right panel of Fig. 4 shows the proportion of List 4 correct recall. Given the success of our PI manipulation, one should expect that the forward testing effect would be substantially weaker (and perhaps altogether absent) in the Low-PI condition than in the High-PI condition. However, similar to Experiment 1, the interaction between PI level and interim activity was again not significant, $F(1,104) = 0.82, p = .367, \eta^2_p = .01, B_{01} = 2.82$. Replicating the results of Experiment 1, reducing PI did not diminish the forward testing effect. Specifically, the tested participants outperformed the restudied participants in both the Low-PI condition ($M_{test} = .68, M_{restudy} = .45$), $t(51) = 3.98, p < .001, d = 1.09, B_{10} = 111.69$, and the High-PI condition ($M_{test} = .61, M_{restudy} = .30$), $t(53) = 4.68, p < .001, d = 1.20, B_{10} = 894.56$. The powerful forward testing effect in the Low-PI condition is especially notable given that participants in this condition produced hardly any intrusions at all. Despite our PI manipulation having little influence on the forward testing effect, there was once again a strong negative correlation between correct recall and intrusions (see Fig. 5), $r(106) = -.44, p < .001, B_{10} = 8395.88$, and this effect was again driven by participants in the restudy condition, $r(48) = -.41, p = .003, B_{01} = 12.06$, rather than those in the test condition, $r(56) = -.14, p = .296, B_{01} = 3.58$.

Discussion

In this experiment, we attempted to further reduce PI for participants in the Low-PI condition. To this end, we had the Low-PI group study words from a single category for all lists, which reduced intrusions to near floor levels for participants in the Low-PI condition. Despite this result, the forward testing effect remained robust, and its magnitude was unaffected by the PI manipulation.

Taken together, the findings of our experiments clearly differed from those reported by Nunes and Weinstein (2012). As stated in the Introduction, the combination of their Experiments 1 and 2 is comparable to our High-PI and Low-PI conditions. Specifically, participants in their Experiment 1 studied several DRM themes in each list, which is similar to our High-PI condition, and in
their Experiment 2, participants studied individual DRM themes in each list, which corresponds to our Low-PI condition. A forward testing effect was found in their Experiment 1, but the effect was entirely absent in their Experiment 2.

Why do the current study and Nunes and Weinstein (2012) show contrasting results? Although the design of our experiments was modeled after Nunes and Weinstein, several methodological differences remained. Perhaps the most obvious was the materials. Namely, we used categorized lists, whereas Nunes and Weinstein used DRM lists, with the former constructed based on category norms and the latter constructed based on associative norms. Another methodological difference between the studies was the comparison task. In both studies, all participants completed 1 min of math problems after studying each word list. Following this, participants in the control condition restudied the words in our experiments, whereas they solved additional math problems in Nunes and Weinstein’s experiments.

Despite these methodological differences, we could not conceive of a compelling explanation that would account for the vastly different results. In particular, it is difficult to envision these methodological differences having different effects on the magnitude of the forward testing effect depending on whether participants are expected to experience higher or lower PI. To our knowledge, no existing theoretical explanations (Chan, Meissner et al., 2018) would lead one to predict that a manipulation of PI would affect the forward testing effect with associative DRM words but not with categorized words (or with using math problems as the control task but not with restudy as the control task). Consequently, we aimed to directly replicate Nunes and Weinstein’s Experiment 2.

**Experiment 3**

The goal of Experiment 3 was to faithfully replicate Experiment 2 of Nunes and Weinstein (2012). To this end, we contacted the primary author of that study and confirmed that our materials and procedure represent a faithful, direct replication of their study (Nunes, personal communication, 2020). There were four notable differences between Experiment 3 and our Experiments 1 and 2. First, participants studied DRM words instead of categorized words. Second, participants studied 12 words per list instead of 16 words per list. Third, each word was shown for 2 s and only once (in our Experiments 1 and 2, participants studied each word twice at 4 s apiece). Lastly, participants in the control condition solved math problems instead of restudying the list.

**Method**

**Design and participants**

Mirroring Nunes and Weinstein (2012), Experiment 3 used a one-factor, two-level design (i.e., test vs. no-test), with this factor manipulated between subjects. A total of 71 undergraduate students from Iowa State University participated for course credit. Data from four participants were not analyzed – two because they did not follow instructions and two because of an experimenter error. The final sample thus included data from 67 participants, with 36 in the test condition and 31 in the no-test condition. Sensitivity power analysis indicated that this sample size had .80 power to detect an effect size of $d = 0.70$.

**Materials and procedure**

The materials and procedure were the same as those of Nunes and Weinstein (2012), Experiment 2. Four DRM lists (mountain, soft, river, and city) from the Stadler et al. (1999) norms were used as study material. Lists 1–4 contained 12 words from a single DRM theme. Words in the third, fifth, and 13th positions of each DRM theme in the Stadler norms were presented only in the target List 5. Four different counterbalances were used, such that some participants studied the soft list as List 1 and others studied it as their List 2, 3, or 4. However, the order of words within each list was fixed and presented in the same order as they appeared in the Stadler et al. norms (i.e., hill was always presented first, then valley, then summit, etc.). Note that the counterbalancing scheme affected only words in Lists 1–4; the words selected for List 5 were not counterbalanced, such that these words never appeared in Lists 1–4 for any participants. List 5 presented the three previously non-studied words from each studied DRM theme in a different random order for each participant, but all participants studied the same words in List 5.

All words were presented for 2 s each with a 500-ms inter-stimulus interval. After each list, participants solved ten math problems at 6 s each. Then, participants in the test condition had 60 s to recall the words from the just-studied list, whereas those in the no-test condition solved additional math problems for another 60 s. After studying List 5, all participants solved math problems for 60 s and then took a recall test. They then completed a 5-min cumulative recall test. The cumulative test was conducted as part of the replication, but data from this test were not of interest and were therefore not scored.

**Results**

**Intrusions during List 5 recall**

In Experiments 1 and 2, we used frequency to report intrusion data, which is the predominant metric in the literature
(Chan, Manley et al., 2018; Weinstein et al., 2014). But here, we report intrusions using frequency as well as proportions because the latter was used to present intrusions in Nunes and Weinstein (2012). Specifically, they divided the frequency of intrusions by 60, which was the number of study words across all five lists. Consistent with Nunes and Weinstein, we did not count false recall of the critical DRM words as intrusions. Instead, these words were scored separately as critical false recall.

The outlier cutoff for intrusions was 4.03 items ($M = 0.97$, $SD = 1.53$), which excluded data from four participants. The left panel of Fig. 6 shows the frequency of intrusions during List 5 recall. Interestingly, unlike the results in Experiments 1 and 2, there was no significant difference in intrusions between the tested participants ($M = 0.50$, $M = .01$ in proportions) and the nontested participants ($M = 0.89$, $M = .02$ in proportions), $t(61) = 1.66$, $p = .102$, $d = 0.42$, $B_{10} = 1.22$. These data are comparable to the findings of Nunes and Weinstein (2012), who found no significant effect of intervening activity on intrusions. Interestingly, when we examined the intrusion data with the outliers included, interpolated testing significantly reduced intrusions ($d = 0.72$, see OSM for details), thereby replicating the pattern reported in Experiments 1 and 2.

Following Nunes and Weinstein (2012), critical false recall rates were reported as proportions (with the frequency of critical false recall divided by 4 because there were four DRM themes). Similar to Nunes and Weinstein’s results ($M_{test} = 0.04$, $M_{no-test} = 0.06$), the recall rate of critical words did not differ significantly between the test and no-test conditions ($M_{test} = 0.08$, $M_{no-test} = 0.15$), $t(61) = 0.70$, $p = .489$, $d = 0.18$, $B_{01} = 3.14$.

**List 5 correct recall**

The most important results are presented in the right panel of Fig. 6. Similar to our Experiments 1 and 2 but in contrast to Nunes and Weinstein (2012), we observed a robust forward testing effect on List 5 correct recall, $t(61) = 2.93$, $p = .005$, $d = 0.75$, $B_{10} = 8.39$ ($M_{test} = .44$, $M_{no-test} = .29$). This finding is consistent with what we reported in Experiments 1 and 2, but differed from Nunes and Weinstein (2012), who found no forward testing effect ($d \approx 0$, $M_{test} = .46$, $M_{no-test} = .46$). Moreover, replicating results of our Experiments 1 and 2, a negative correlation between intrusions and correct recall was observed (see Fig. 7), $r(61) = -.35$, $p = .005$, $B_{10} = 7.32$. Unlike Experiments 1 and 2, the correlation was contributed by participants in both the test condition, $r(34) = -.34$, $p = .042$, $B_{01} = 1.51$, and the no-test condition, $r(25) = -.27$, $p = .176$, $B_{01} = 1.75$. We believe that the nonsignificant correlation in the no-test condition might be due to insufficient power.

**Discussion**

The results in Experiment 3 largely resembled those from our Experiments 1 and 2 – that is, interpolated testing facilitated new learning even when participants studied words in an environment that was meant to minimize the buildup of PI. This finding contrasts directly with that reported by Nunes and Weinstein (2012), and is particularly intriguing when viewed in the context of the overall replication effort. Specifically, although our study used a different participant sample and the data were collected nearly a decade later than the original study, in some ways, our data replicated those of the original study exceptionally well. Indeed, most of the
dependent variables, including the amount of studied word intrusions (in proportions, \( M_{\text{test}} = .01, M_{\text{no-test}} = .02 \) in our experiment, \( M_{\text{tested}} = .01, M_{\text{no-test}} = .04 \) in Nunes and Weinstein), false recall of critical words (\( M_{\text{test}} = .02, M_{\text{no-test}} = .04 \) in our experiment, \( M_{\text{tested}} = .04, M_{\text{no-test}} = .06 \) in Nunes and Weinstein), and correct recall of the studied words in the test condition (\( M = .44 \) in our experiment, \( M = .46 \) in Nunes and Weinstein), produced data that were remarkably similar to those reported by Nunes and Weinstein. The lone exception was the correct recall data in the no-test condition. Whereas our participants recalled .29 of the List 5 words, Nunes and Weinstein’s participants recalled .46. Because of the larger sample size of our experiment and its replication of Experiments 1 and 2, we believe that our result probably provided a closer approximation of the population mean. However, given the theoretical weight of this result, we aimed to replicate the experiment again in Experiment 4 but with a full counterbalancing scheme.

**Experiment 4**

**Method**

**Design, participants, materials, and procedure**

A total of 70 undergraduate students from Iowa State University participated for course credit. Unlike Experiments 1–3, in which the data were collected in our laboratory, the data from Experiment 4 were collected online because of the COVID-19 related shutdown of in-person research. Data from three participants (i.e., one in the test condition, two in the no-test condition) were not analyzed because they did not follow instructions. The final sample included data from 67 participants, with 34 in the test condition and 33 in the no-test condition. Sensitivity power analysis showed that this sample size had .80 power to detect an effect size of \( d = 0.69 \).

The materials and procedures were identical to those in Experiment 3 with the following exceptions. In Experiment 3, List 5 always contained the same words. In the current experiment, we fully counterbalanced word assignments to all five lists, such that the words that appeared in List 5 for one participant would appear in List 1, 2, 3, or 4 for another participant. We split each set of 15 DRM theme words into five groups, with each group having similar backward associative strength (BAS), which indicates the strength of the association from the list item to the critical word (Roediger et al., 2001). The five groups of DRM words rotated across the lists across participants. This counterbalancing scheme eliminated the possibility that the results would only be applicable to the particular DRM words chosen to be shown in List 5 of our Experiment 3 and Nunes and Weinstein’s Experiment 2 (2012). Second, the words within every list (and not simply List 5) were presented in a random order. We opted for this procedure because we wanted to avoid having only the target list being presented in a random order. We suspect that Nunes and Weinstein presented their nontarget lists in descending order of associative strength because they wanted to elicit more false memories, which was not important for present purposes. Note that despite these changes, we still consider Experiment 4 a direct replication (Simons, 2014; Zwaan et al., 2017).

**Results**

**Intrusions during List 5 recall**

The outlier cutoff for this experiment was 8.43 intrusions (\( M = 2.01, SD = 3.21 \)), which excluded data from four participants. As can be seen in the left panel of Fig. 8, the tested participants exhibited fewer intrusions (\( M = 0.77, M = .01 \) in proportions) than the nontested participants (\( M = 2.10, M = .04 \) in proportions), \( t(61) = 3.56, p < .001, d = 0.90, B_{10} = 40.74 \). This finding is consistent with our data from Experiments 1 and 2 but differs from our Experiment 3 and Nunes and Weinstein (2012). The critical false recall data also differed from Experiment 3 and Nunes and Weinstein. Specifically, interpolated testing reduced critical false recall (\( M_{\text{test}} = 0.03, M_{\text{no-test}} = 0.10 \), \( t(61) = 2.51, p = .015, d = 0.64, B_{10} = 3.46 \). It is possible that the lack of counterbalancing in our Experiment 3 and Nunes and Weinstein contributed to the null effect of testing on intrusions and critical false recall.

![Fig. 7](image-url)
List 5 correct recall

The right panel of Fig. 8 shows the mean proportion of List 5 correct recall. Mirroring the results from Experiment 3, participants in the test condition ($M = .51$) showed vastly superior performance relative to those in the no-test condition ($M = .17$), $t(61) = 5.60$, $p < .001$, $d = 1.42$, $B_{10} = 23280.95$. Indeed, this is a very large effect and it provides exceptionally strong evidence in favor of the alternative hypothesis.

Once again, there was a negative correlation between intrusions and correct recall (see Fig. 9), $r(61) = -.36$, $p = .004$, $B_{10} = 8.36$. When we split the data by condition, the correlation was not significant in either the no-test condition, $r(27) = -.27$, $p = .160$, $B_{01} = 1.69$, or the test condition, $r(27) = -.02$, $p = .905$, $B_{01} = 4.66$. But the pattern was similar to Experiments 1 and 2, with higher correlation being displayed in the control condition, and insufficient power is likely the cause of these nonsignificant results.

Combined analyses

In the following, we report a meta-analyses that combined the data across our four experiments and the two experiments in Nunes and Weinstein (2012). We conducted a fixed-effects meta-analysis based upon the target list correct recall performance. We coded each High-PI and Low-PI condition as a separate sample, and the overall meta-analytic effect size for the forward test effect across these samples was $d = 0.89$, $Z = 4.63$, $p < .001$. A moderator analysis with the level of PI as the variable of interest showed that differences in PI did not affect the magnitude of the forward testing effect, $Z = 0.41$, $p = .680$, with the High-PI ($d = 0.95$) and Low-PI ($d = .83$) samples producing comparable effect sizes. In fact, the slightly smaller effect size for the Low-PI studies was driven primarily by the inclusion of Nunes and Weinstein’s (2012) Experiment 2 data. When this study was removed from the sample, the mean forward testing effect for the Low-PI studies ($d = 1.01$) was just as large as that for the High-PI studies. An examination of the forest plot in Fig. 10 shows that the data point from Nunes and Weinstein’s Experiment 2 lies outside the confidence interval of the meta-analytic effect size. Therefore, the absence of a forward testing effect in their Experiment 2 could be due to a sampling error and/or a lack of counterbalancing of the target list items.
General discussion

The goal of the present study was to provide a critical test of the release from proactive interference account of the forward testing effect. In Experiments 1 and 2, we used the content change method to manipulate PI and examined whether the size of the forward testing effect varied by PI level. In Experiments 3 and 4, we attempted a direct replication of Nunes and Weinstein’s (2012) study, which largely matched our Low-PI condition in Experiments 1 and 2. If interpolated testing promotes new learning by reducing PI, the forward testing effect should be reduced or eliminated when PI level is low, because there would be little to no PI to be further reduced by testing. On the contrary, across four experiments, we found that interpolated testing consistently promoted new learning regardless of the level of PI that participants experienced.

These results suggest that although interpolated testing can release participants from the negative effects of proactive interference, this benefit is unlikely to be the cause of the forward testing effect. The invariance of the forward testing effect against our PI manipulation might seem paradoxical given that we consistently observed a negative correlation between correct recall and intrusions across participants. At first glance, this latter finding seems to support the release-from-PI account because one interpretation is that interpolated testing reduces response competition during recall, which makes it easier for the tested participants to retrieve the studied items. Our data revealed two problems with this interpretation. First, the negative correlation was mainly driven by the nontested participants. Second and more importantly, the fact that a reduction in intrusions and an increase in correct recall tend to co-occur does not imply causation. In a similar vein, one of the strongest correlates of the (backward) testing effect is retrieval practice performance. In a meta-analysis of the literature, Rowland (2014) showed that studies that reported better retrieval practice performance tended to produce a greater testing effect. Drawing a causal argument based on this association might lead one to predict that retrieval practice would only benefit learning when it is successful, but studies have now shown that unsuccessful retrieval practice can enhance retention (Carneiro et al., 2018; Davis et al., 2017; Knight et al., 2012; Pyc & Rawson, 2010, 2012).

Likewise, the forward benefit of testing appears to be associated with, but not dependent upon, the release-from-PI effect. We believe that the negative correlation between intrusions and correct recall is driven by a third variable, and one possible candidate is that testing promotes semantic organization in memory (Chan, Manley et al., 2018; Zaromb & Roediger, 2010). Recent research has shown that, across successive tests, learners become increasingly likely to cluster semantically related items during recall, and the level of clustering is positively associated with correct recall performance (Chan et al., 2020; Yang et al., 2021). We have recently attributed this improved memory organization to participants optimizing their encoding and retrieval strategies across lists (Chan et al., 2020; Chan, Manley et al., 2018). Although studies in the forward testing literature have only investigated clustering based on semantic relationships, other studies have shown that testing can also increase temporal clustering and segregation. Specifically, interpolated testing can promote learners’ ability to cluster items that were studied in close temporal proximity (i.e., items studied in the same list), while simultaneously segregating items studied pre- or post-retrieval (i.e., items studied across lists; Brewer et al., 2010; Chan & McDermott, 2007; Jang & Huber, 2008; Kliegl & Bäuml, 2021; Wahlheim, 2015; Yang et al., 2021). As Darley and Murdock (1971) speculated, switching between an encoding and retrieval task might provide an opportunity for learners to mentally label or tag items based on whether they were previously tested or not (Davis et al., 2017), and this labeling mechanism can reduce intrusions (Pierce et al., 2017) either through an early selection or late correction mechanism. From the perspective of early selection, participants might constrain their retrieval at the front end by attempting their memory search through items that they had not previously retrieved (Rugg et al., 2000; Shimizu & Jacoby, 2005); alternatively, participants might attempt a broad memory search and then withhold output of previously recalled items. The early selection mechanism – that is, constrained retrieval – might serve to both increase correct recall (by reducing the search set) and reduce intrusions (because previously tested items are not in the search set), thereby producing the negative correlation between these dependent variables.
Recently, researchers have claimed that the PI-reducing effect of retrieval contributes to the forward testing effect (Dang et al., 2021; Yang et al., 2021). Support for this argument was based upon a mediation analysis in which prior list intrusions served as a significant mediator between the independent variable of interpolated activity (i.e., testing vs. restudy) and the dependent variable of correct recall, and the researchers interpreted this indirect effect as retrieval insulating learners from PI, which in turn promotes new learning. However, we caution against such an interpretation in the current case because there was no temporal precedence between the mediator (i.e., intrusions) and the dependent variable (i.e., correct recall). Instead, intrusions and correct recall occur during the same experimental phase, making it impossible to infer causality via temporal orders (i.e., something that occurs later cannot cause something that occurs earlier; Aalen et al., 2012; Hayes, 2018; Preacher, 2015; Walters, 2019). Indeed, it is possible to reverse the roles of the mediator and the dependent variable in the analyses performed by Dang et al. (2021) and Yang et al. (2021), such that correct recall serves as the mediator and intrusion serves as the dependent variable. To explore this possibility, we conducted a mediation analysis using data that Dang et al. and Yang et al. provided on the Open Science Framework.

We first attempted to replicate their original mediation analyses, and the results of our model were consistent with theirs, such that there was a significant indirect effect of intrusions on correct recall. We then examined the viability of the reversed mediation model. As we had suspected, the indirect effect of this reversed model was also significant, -0.238 [-0.418, -0.112] in Dang et al., and -0.942 [-1.217, -0.666] in Yang et al., indicating that interpolated testing affects prior list intrusions through correct recall. The significance of this reversed model highlights the difficulty when interpreting the results of a mediation analysis in which the mediator and dependent variable have no clear temporal precedence.

We would be remiss not to mention that in a recent report, Chan, Manley, et al. (2018) suggested that interpolated testing might improve new learning in the categorized lists learning paradigm by increasing the likelihood that participants would notice the categorized nature of the lists. The idea is that once participants realize the categorized nature of the lists, they would attempt to encode the semantic relationship across similar list items, which should enhance learning. Chan et al. further reasoned that, if this hypothesis is correct, making the categorized nature of the list readily discoverable (e.g., by presenting words from the same category in blocks) should reduce the forward testing advantage, and they cited Nunes and Weinstein (2012) as providing preliminary support for such a hypothesis. The present results are clearly at odds with these predictions.

Limitations

One limitation of this study is that only the retrieval form of PI was addressed because we measured intrusions. But encoding can also be affected by the buildup of PI, because learners’ attentional resources can deteriorate as they continue to encode information. For example, in an EEG study, Pastötter and colleagues (2011) reported that interpolated testing enhanced the encoding of new information by reducing inattention, as evidenced by an increase in alpha power (an electrophysiological correlate of inattention) across lists for the control participants, but not for the tested participants. Furthermore, Szpunar and colleagues (Szpunar et al., 2013; Jing et al., 2016; see also Pan et al., 2020, for a similar finding with pretesting rather than interpolated testing) showed that, when viewing a video-recorded lecture, participants who received an interpolated test every 5 min were much less likely to report mind wandering than control participants. However, as stated earlier, we believe that such attention-based mechanisms should be treated as distinct from the release-from-PI account, because allowing this account to include attentional/encoding-based mechanisms would make the account too broad and difficult to falsify.

Our studies, like many behavioral studies, also did not allow us to distinguish encoding- and retrieval-based effects, because our encoding manipulation (i.e., High-PI vs. Low-PI) could, and likely did, have a cascading effect on retrieval. Future studies might attempt to tease apart encoding and retrieval effects of the forward testing effect using brain imaging or electrophysiological methods that provide real-time data during encoding (Pastötter et al., 2011; but see also Beardsley et al., 2018).

Another issue we want to address is the operational definition of PI. In the present experiments, we have used frequency of intrusions as a measure of PI, as did most studies in the forward testing effect literature. But in the verbal learning era, correct recall performance was regularly used to demonstrate PI effects. For example, when comparing memory performance between participants who study a single A-D list (with no prior study) against participants who studied an A-B list before the A-D list, a reduction in A-D recall due to prior learning was considered as evidence for

---

2 We conducted our analysis in JASP, whereas Dang et al. and Yang et al. conducted their analyses in R. For Dang et al., we used the data from adult population and the coefficient of the indirect effect of intrusions on critical recall was 0.072 [0.037, 0.124]. Although the exact coefficient was different from the original paper (0.093 [0.061, 0.129]), it was in the same direction and the difference might be attributed to the different assumptions underlying the different statistical packages. For Yang et al., the coefficients were identical, 1.199 [0.856, 1.542] in our replication, and 1.199 [0.902, 1.503] in the original paper.
PI. Adopting this definition, one might argue that our PI manipulation had no impact because the control participants consistently demonstrated poorer correct recall performance on the target list relative to the List 1 recall performance of the tested participants (which is equivalent to a group of participants who had no prior study, see Table 1). However, adopting this definition of PI for the forward testing effect would render the release-from-PI account purely descriptive (rather than explanatory), because whenever a forward testing effect is found, by definition, participants in the tested condition must outperform those in the control condition. Claiming such a difference as release-from-PI simply restates, rather than explains, the phenomenon.

Lastly, we acknowledge that our sample size was too small to detect an interaction between interim activity and PI level. If a moderate effect size (e.g., $f = .2$) was assumed for the interaction effect with an alpha level of .05 and .80 power, a total sample size of 199 participants would have been required. One might notice that the forward testing effect size was smaller in the Low-PI condition than in the High-PI condition from both Experiment 1 (Low-PI = .14, High-PI = .20) and Experiment 2 (Low-PI = .23, High-PI = .31), and one interpretation of these nonsignificant differences is that the forward testing effect was less pronounced when participants experienced less PI. However, we caution against over-interpreting differences based on raw probabilities because they do not account for differences in data variability. When we examine the same data with Cohen’s $d$, it becomes clear that the differences in percentage are more apparent than real. Specifically, the forward testing effect is .81 in the Low-PI condition and .82 in the High-PI condition in Experiment 1, and 1.09 in the Low-PI condition and 1.20 in the High-PI condition in Experiment 2. Nonetheless, it remains possible that the interaction between PI condition and interim activity was not significant due to low statistical power. In other words, release-from-PI might have made a small contribution to the forward testing effect, but our experiments were not sufficiently powered to detect it. Thus, future research investigating the relationship between PI level and interim activity would benefit from a larger sample.

**Concluding remarks**

In the present experiments, we provided a critical test of the release-from-PI account, which is widely considered to be one of the key candidate explanations for the forward testing effect. Across four experiments, we consistently observed data that violate its predictions. Importantly, we failed to replicate Nunes and Weinstein’s (2012) finding, which provided heretofore the crucial piece of evidence for the release-from-PI account. At the very least, our data suggest that, should release-from-PI contribute to the forward testing effect, it likely plays a smaller role than previously expected.

Because our study was designed to test the release-from-PI hypothesis, it could not address the question of “if release from proactive interference is not a causal factor for the forward testing effect, what is?” Fortunately, several potentially viable explanations remain. For example, interpolated testing might potentiate new learning by promoting attentive encoding (Chan et al., 2020; Szpunar, 2017; Szpunar et al., 2013; Weinstein et al., 2014). Another possibility is that testing might enable learners to develop more effective encoding/retrieval strategies that take advantage of better memory organization (Chan et al., 2020; Dang et al., 2021) or contextual segregation (Yang et al., 2021). Future research that directly tackles these accounts will help promote our understanding of the forward testing effect.

**Appendix 1**

**Experiment 1**

| Category | Fruits | Animals | Body parts | Sports |
|----------|--------|---------|------------|--------|
| Presented Items | grape | lion | hand | tennis |
| | strawberry | bear | mouth | hockey |
| | kiwi | cow | heart | swimming |
| | pineapple | elephant | knee | golf |
| | watermelon | deer | neck | volleyball |
| | papaya | rat | brain | lacrosse |
| | plum | giraffe | elbow | track |
| | mango | squirrel | shoulder | rugby |
| | cherry | rabbit | chest | softball |
| | lemon | goat | thigh | skiing |
| | cantaloupe | zebra | liver | running |
| | lime | moose | lung | gymnastics |
| | nectarine | sheep | wrist | polo |
| | apricot | wolf | teeth | racquetball |
| | guava | turtle | muscle | wrestling |
| | coconut | frog | abdomen | bowling |
| | pear | tiger | nose | badminton |
| | blueberry | pig | hair | cricket |
| | pomegranate | lizard | nail | snowboarding |
| | honeydew | elk | spine | cheerleading |
**Experiment 2**

| Category         | Fruits          | Animals        | Body parts     | Furniture      |
|------------------|-----------------|----------------|----------------|----------------|
| Presented Items  | grape           | lion           | hand           | bed            |
|                  | strawberry      | bear           | mouth          | desk           |
|                  | kiwi            | bow            | heart          | sofa           |
|                  | pineapple       | elephant       | knee           | dresser        |
|                  | watermelon      | deer           | neck           | loveseat       |
|                  | papaya          | rat            | brain          | lamp           |
|                  | plum            | giraffe        | elbow          | nightstand     |
|                  | mango           | squirrel       | shoulder       | ottoman        |
|                  | cherry          | rabbit         | chest          | recliner       |
|                  | lemon           | goat           | thigh          | stool          |
|                  | cantaloupe      | zebra          | liver          | futon          |
|                  | lime            | moose          | lung           | armoire        |
|                  | nectarine       | sheep          | wrist          | cabinet        |
|                  | apricot         | wolf           | teeth          | bookshelf      |
|                  | guava           | turtle         | muscle         | rug            |
|                  | coconut         | frog           | abdomen        | closet         |

**Experiment 3 and 4**

| DRM theme       | City            | Mountain       | River           | Soft         |
|-----------------|-----------------|----------------|-----------------|--------------|
| Presented items | metropolis      | climber        | Mississippi     | hard         |
|                 | big             | top            | water           | light        |
|                 | village         | plain          | boat            | kitten       |
|                 | town            | hill           | creek           | loud         |
|                 | subway          | glacier        | tide            | feather      |
|                 | crowded         | range          | swim            | skin         |
|                 | New York        | climb          | stream          | tender       |
|                 | country         | ski            | run             | cotton       |
|                 | state           | steep          | barge           | fur          |
|                 | urban           | molehill       | flow            | fluffy       |
|                 | streets         | summit         | lake            | plush        |
|                 | Chicago         | bike           | winding         | touch        |
|                 | suburb          | peak           | bridge          | pillow       |
|                 | county          | valley         | brook           | downy        |
|                 | capital         | goat           | fish            | furry        |

References

Aalen, O. O., Raysland, K., Gran, J. M., & Ledergerber, B. (2012). Causality, mediation and time: A dynamic viewpoint. *Journal of the Royal Statistical Society. Series A: Statistics in Society, 175*(4), 831–861. https://doi.org/10.1111/j.1467-985X.2011.01030.x

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

Beardsley, M., Hernández-Leo, D., & Ramírez-Melendez, R. (2018). Seeking reproducibility: Assessing a multimodal study of the testing effect. *Journal of Computer Assisted Learning, 34*(4), 378–386. https://doi.org/10.1111/jcal.12265

Bonner, S. M. (2009). Investigating teacher use of practice tests for formative purposes. *Journal of Multidisciplinary Evaluation, 6*(12), 125–136.

Brewer, G. A., Marsh, R. L., Meeks, J. T., Clark-Foos, A., & Hicks, J. L. (2010). The effects of free recall testing on subsequent source memory. *Memory, 18*(4), 385–393. https://doi.org/10.1080/09658211003702163

Brown, J. (1958). Some tests of the decay theory of immediate memory. *Quarterly Journal of Experimental Psychology, 10*(1), 12–21. https://doi.org/10.1080/17470218508416249

Carneiro, P., Lapa, A., & Finn, B. (2018). The effect of unsuccessful retrieval on children’s subsequent learning. *Journal of Experimental Child Psychology, 166*, 400–420. https://doi.org/10.1016/j.jecp.2017.09.010

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

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*(3), 213–229. https://doi.org/10.1016/j.jarmac.2017.07.003

Chan, J. C. K., Manley, K. D., Davis, S. D., & Szpunar, K. K. (2018). Testing potentiates new learning across a retention interval and a lag: A strategy change perspective. *Journal of Memory and Language, 102*(May), 83–96. https://doi.org/10.1016/j.jml.2018.05.007

Chan, J. C. K., Meissner, C. A., & Davis, S. D. (2018). Retrieval potentiates new learning: A theoretical and meta-analytic review. *Psychological Bulletin, 144*(11), 1111–1146. https://doi.org/10.1037/bul0000166

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

Conway, A. R. A., Kane, M. J., Bunting, M. F., Hambrick, D. Z., Wilhelm, O., & Engle, R. W. (2005). Working memory span tasks: A methodological review and user’s guide. *Psychonomic Bulletin and Review, 12*(5), 769–786. https://doi.org/10.1012/jbcr.200512.001

Crowder, R. G. (1976). *Principles of learning and memory.* Erlbaum.

Dang, X., Yang, C., & Chen, Y. (2021). Age difference in the forward testing effect: The roles of strategy change and release from proactive interference. *Cognitive Development, 59*(June), 101079. https://doi.org/10.1016/j.cogdev.2021.101079

Darley, C. F., & Murdock, B. B. (1971). Effects of prior free recall testing on final recall and recognition. *Journal of Experimental Psychology, 91*, 66–73.

Davis, S. D., & Chan, J. C. K. (2015). Studying on borrowed time: How does testing impair new learning? *Journal of Experimental Psychology: Learning, Memory, and Cognition, 41*(6), 1741–1754. https://doi.org/10.1037/xmr0000166

Davis, S. D., Chan, J. C., & 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, 6*(4), 434-441. https://doi.org/10.1016/j.jarmac.2017.07.002
Engle, R. (1975). Pupillary measurement and release from proactive inhibition. *Perceptual and motor skills, 41*(3), 835–842. https://doi.org/10.2466/pms.1975.41.3.835

Faul, F., Erdfelder, E., Lang, A. G., & Buchner, A. (2007). A Cal- orimetric investigation of the liquid bi-ni alloys. *Behavior Research Methods, 39*(2), 175–191. https://doi.org/10.3758/BF03193146

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. https://doi.org/10.1037/a0032377

Gardiner, J. M., Craik, F. I., & Birtwistle, J. (1972). Retrieval cues and release from proactive inhibition. *Journal of Verbal Learning and Verbal Behavior, 11*(6), 778–783. https://doi.org/10.1016/S0022-5371(72)80012-4

Hayes, A. F. (2018). Causal Steps, Confounding, and Causal Order. In A. F. Hayes (Ed.), *Introduction to mediation, moderation, and conditional process analysis: a regression based approach* (2nd ed., pp. 113–146). Guiford Press.

Jang, Y., & Huber, D. E. (2008). Context retrieval and context change in free recall: Recalling from long-term memory drives list isolation. *Journal of Experimental Psychology: Learning Memory and Cognition, 34*(1), 112–127. https://doi.org/10.1037/0278-7393.34.1.112

JASP Team (2020). JASP (Version 0.14)[Computer software].

Jing, H. G., Szpunar, K. K., & Schacter, D. L. (2016). Interpolated testing influences focused attention and improves integration of information during a video-recorded lecture. *Journal of Experimental Psychology: Applied, 22*(3), 305–318. https://doi.org/10.1037/xap0000087

Karpicke, J. D., & Roediger, H. L. (2008). The critical importance of retrieval for learning. *Science, 319*(5865), 966–968. https://doi.org/10.1126/science.1152408

Karpicke, J. D., Butler, A. C., & Roediger, H. L. (2009). Metacognitive strategies in student learning: Do students practise retrieval when they study on their own? *Memory, 17*(4), 471–479. https://doi.org/10.1080/09658210902647009

Kliegl, O., & Bäuml, K. H. T. (2021). Buildup and release from proactive interference – Cognitive and neural mechanisms. *Neuroscience and Biobehavioral Reviews, 120*, 264–278. https://doi.org/10.1016/j.neubiorev.2020.10.028

Knight, J. B., Hunter Ball, B., Brewer, G. A., DeWitt, M. R., & Marsh, R. L. (2012). Testing unsuccessfully: A specification of the underlying mechanisms supporting its influence on retention. *Journal of Memory and Language, 66*(4), 731–746. https://doi.org/10.1016/j.jml.2011.12.008

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. https://doi.org/10.1080/09658211.2011.648198.

Pan, S. C., Sana, F., Schmitt, A. G., & Bjork, E. L. (2020). Pretesting reduces mind wandering and enhances learning during online lectures. *Journal of Applied Research in Memory and Cognition, 9*(4), 542–554. https://doi.org/10.1016/j.jarmac.2020.07.004

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

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

Peterson, L., & Peterson, M. J. (1959). Short-term retention of individual verbal items. *Journal of Experimental Psychology, 58*(3), 193–198. https://doi.org/10.1037/h0049234

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, 43*(7), 1063–1072. https://doi.org/10.1037/xlm0000377

Postman, L., & Hasher, L. (1972). Conditions of proactive inhibition in free recall. *Journal of Experimental Psychology, 92*(2), 276–284. https://doi.org/10.1037/h00352074

Postman, L., & Keppel, G. (1977). Conditions of cumulative proactive inhibition. *Journal of Experimental Psychology: General, 106*(4), 376–403. https://doi.org/10.1037/0096-3445.106.4.376

Preacher, K. J. (2015). Advances in mediation analysis: A survey and synthesis of new developments. *Annual Review of Psychology, 66*, 825–852. https://doi.org/10.1146/annurev-psych-010814-015258

Pyc, M. A., & Rawson, K. A. (2010). Why testing improves memory: Mediator effectiveness hypothesis. *Science, 330*(6002), 335. https://doi.org/10.1126/science.1191465

Pyc, M. A., & Rawson, K. A. (2012). Why is test-restudy practice beneficial for memory? An evaluation of the mediator shift hypothesis. *Journal of Experimental Psychology: Learning Memory and Cognition, 38*(3), 737–746. https://doi.org/10.1037/a0026166

Roediger, H. L., Watson, J. M., McDermott, K. B., & Gallo, D. A. (2003). Factors that determine false recall: A multiple regression analysis. *Psychonomic Bulletin & Review, 8*, 385–407. https://doi.org/10.3758/BF03196177

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. https://doi.org/10.1037/0278-7393.21.4.803

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. https://doi.org/10.1037/a0037559

Rugg, M. D., Allan, K., & Birch, C. S. (2000). Electrophysiological evidence for the modulation of retrieval orientation by depth of study processing. *Journal of Cognitive Neuroscience, 12*(4), 664–678. https://doi.org/10.1162/089892900562291

Postman, L., Stark, K., & Fraser, J. (1968). Temporal changes in interference. *Journal of Verbal Learning and Verbal Behavior, 7*(3), 672–694. https://doi.org/10.1016/S0022-5371(68)80124-0

Shimizu, Y., & Jacoby, L. L. (2005). Similarity-guided depth of retrieval: Constraining at the front end. *Canadian Journal of Experimental Psychology, 59*(1), 17–21. https://doi.org/10.1037/ 00087455

Simons, D. J. (2014). The value of direct replication. *Perspectives on Psychological Science, 9*(1), 76–80. https://doi.org/10.1177/174591613514755

Stadler, M. A., Roediger, H. L., & McDermott, K. B. (1999). Norms for word lists that create false memories. *Memory & Cognition, 27*(3), 494–500. https://doi.org/10.3758/BF03211543

Szpunar, K. K. (2017). Directing the wandering mind. *Current Directions in Psychological Science, 26*(1), 40–44. https://doi.org/10.1177/0963721416670320

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. https://doi.org/10.1037/a0013082

Szpunar, K. K., Moulton, S. T., & Schacter, D. L. (2013). Mind wandering and education: From the classroom to online learning. *Frontiers in Psychology, 4*, 1–7. https://doi.org/10.3389/fpsyg.2013.00495

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*(2), 181–193. https://doi.org/10.1016/S0022-5371(74)80043-5
Underwood, B. J. (1975). Individual differences as a crucible in theory construction. *American Psychologist, 30*(2), 128–134. https://doi.org/10.1037/h0076759

Wahlheim, C. N. (2015). Testing can counteract proactive interference by integrating competing information. *Memory and Cognition, 43*(1), 27–38. https://doi.org/10.3758/s13421-014-0455-5

Walters, G. D. (2019). Why are mediation effects so small? *International Journal of Social Research Methodology, 22*(2), 219–232. https://doi.org/10.1080/13645579.2018.1517232

Weinstein, Y., McDermott, K. B., & Chan, J. C. K. (2010). True and false memories in the DRM paradigm on a forced choice test. *Memory, 18*(4), 375–384. https://doi.org/10.1080/09658211003685533

Weinstein, Y., McDermott, K. B., & Szpunar, K. K. (2011). Testing protects against proactive interference in face-name learning. *Psychonomic Bulletin and Review, 18*(3), 518–523. https://doi.org/10.3758/s13423-011-0085-x

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. https://doi.org/10.1037/a0036164

Wickens, D. D., Born, D. G., & Allen, C. K. (1963). Proactive inhibition and item similarity in short-term memory. *Journal of Verbal Learning and Verbal Behavior, 2*(5-6), 440-445. https://doi.org/10.1016/S0022-5371(63)80045-6

Wickens, D. D., Moody, M. J., & Dow, R. (1981). The nature and timing of the retrieval process and of interference effects. *Journal of Experimental Psychology: General, 110*(1), 1–20. https://doi.org/10.1037/0096-3445.110.1.1

Wixted, J., & Rohrer, D. (1993). Proactive interference and the dynamics of free recall. *Journal of Experimental Psychology, 19*(5), 1024–1039.

Yang, C., Potts, R., & Shanks, D. R. (2018). Enhancing learning and retrieval of new information: a review of the forward testing effect. *npj Science of Learning, 3*(1), 1–9. https://doi.org/10.1038/s41539-018-0024-y

Yang, C., Zhao, W., Luo, L., Sun, B., Potts, R., & Shanks, D. R. (2021, July 22). Testing potential mechanisms underlying test-potentiated new learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*. Advance online publication. https://doi.org/10.1037/slm0001021

Zaromb, F. M., & Roediger, H. L. (2010). The testing effect in free recall is associated with enhanced organizational processes. *Memory and Cognition, 38*(8), 995–1008. https://doi.org/10.3758/MC.38.8.995

Zwaan, R. A., Etz, A., Lucas, R. E., & Donnellan, M. B. (2017). Making Replication Mainstream. *Behavioral and Brain Sciences, (2018), 1–50*. https://doi.org/10.1017/S0140525X17001972

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.