Testing Potential Mechanisms Underlying Test-Potentiated New Learning

Chunliang Yang[1], Wenbo, Zhao[2], Liang Luo[1,2], Bukuan Sun[3], Rosalind Potts[4], David R. Shanks[4]

[1] Institute of Developmental Psychology, Faculty of Psychology, Beijing Normal University, Beijing, China.

[2] Collaborative Innovation Center of Assessment Toward Basic Education Quality, Beijing Normal University, Beijing, China.

[3] School of Education, Fuqing Branch of Fujian Normal University, Fuqing, China.

[4] Division of Psychology and Language Sciences, University College London, London, the UK.

Author Note

The data contained in this project are publicly available at Open Science Framework (https://osf.io/76wm8/). Correspondence concerning this article should be addressed to Liang Luo, 19 Xinjiekou Wai Street, Beijing 100875, China. E-mail: luoliang@bnu.edu.cn

Acknowledgments

This research was supported by the Natural Science Foundation of China (32000742; 31671130), the Fundamental Research Funds for the Central Universities (2019NTSS28), and the United Kingdom Economic and Social Research Council (ES/S014616/1).

Abstract

An emerging body of studies demonstrates that practicing retrieval of studied information, by comparison with restudying or no treatment, can facilitate subsequent learning and retrieval of new information, a phenomenon termed the *forward testing effect* (FTE) or *test-potentiated new learning*. Several theoretical explanations have been proposed to account for the FTE. A release-from-PI theory proposes that interpolated testing induces context changes and enhances event segregation, which in turn protect new learning from proactive interference (PI). A strategy-change view hypothesizes that prior tests teach learners to adopt more effective/elaborative learning and retrieval strategies in the subsequent study and test phases. Finally, a reset-of-encoding account proposes that interim testing on studied information reduces memory load, resets the subsequent encoding process, and enhances the encoding of new information. The current study recruited a large sample (over 1,000 participants) and employed a multi-list learning task and mediation analyses to test these theories. The results suggest that prior list intrusions (an index of PI) significantly mediated the FTE, supporting the release-from-PI theory. In addition, interim testing enhanced strategic processing of temporal information during new learning (reflected by increased clustering), and temporal clustering significantly mediated the FTE, supporting a role for strategy-change in the FTE. Lastly, a variety of indices were constructed to represent the benefit of reset-of-encoding, but none of them provided evidence supporting the reset-of-encoding view. The results shed new light on the complex mechanisms underlying the forward benefits of testing.

Keywords: forward testing effect; release-from-PI; temporal processing; strategy-change; reset-of-encoding

Exploring efficient techniques to facilitate learning and consolidate memory has long been a goal of experimental psychology (Ebbinghaus, 1885/1913). One of the most widely studied and also most efficient strategies is testing (i.e., retrieval practice). Over the last century (Abbott, 1909), hundreds of studies both in the laboratory and classroom have repeatedly demonstrated that testing is a more powerful strategy to consolidate long-term retention of studied information than other methods, such as restudying, note-taking, concept mapping, and so on (for reviews, see Roediger & Karpicke, 2006; Rowland, 2014; Yang, Luo, Vadillo, Yu, & Shanks, 2020). In the current study, we term the phenomenon, that testing of studied information enhances its long-term retention, the *backward testing effect* (BTE; following precedents, such as Pastötter & Bäuml, 2014; Yang, Potts, & Shanks, 2017). In addition to this backward benefit, an emerging body of recent studies has found that retrieving studied information, by comparison with restudying or no treatment, can also more effectively potentiate subsequent learning and retention of new information, a phenomenon termed the *forward testing effect* (FTE; Pastötter & Bäuml, 2014; Yang, Potts, & Shanks, 2018) or *test-potentiated new learning* (Chan, Manley, Davis, & Szpunar, 2018). The current research attempts to shed light on the mechanisms underlying the FTE.

Although the FTE was incidentally identified about fifty years ago (Tulving & Watkins, 1974), research interest accelerated after a publication by Szpunar, McDermott, and Roediger (2008). Szpunar et al. instructed two groups (Test/No-Test) of participants to study five 18-word lists, which were studied one-by-one (2 sec each) and list-by-list. Following study of each of Lists 1-4, the Test group took a free recall test (i.e., recalling the words from the just-studied list), whereas the No-Test group instead solved irrelevant arithmetic problems. After studying List 5, both groups recalled as many List 5 words as they could. The results showed that the Test group (*M* = 7.00 out of 18) correctly recalled twice as many List 5 words as the No-Test group (*M* = 3.50), clearly demonstrating the FTE: interim testing on Lists 1-4, by comparison with no testing, substantially potentiated learning and retrieval of List 5 words. In addition, Szpunar et al. (2008) observed that their No-Test group (*M* = 3.70) suffered from over ten times as many prior list intrusions (i.e., incorrectly recalling List 1-4 words when instructed to recall List 5) as their Test group (*M* = 0.30), indicating that, besides potentiating new learning, interpolated testing can also prevent the build-up of proactive interference (PI; i.e., prior learning interfering with subsequent new encoding) across learning events.

Subsequently, dozens of studies have been conducted to explore the FTE on different types of learning and its generalizability to different situations (for reviews, see Chan, Meissner, & Davis, 2018; Pastötter & Bäuml, 2014; Yang et al., 2018). For instance, it has been found that the FTE generalizes across different types of learning, such as the learning of single items (e.g., Szpunar et al., 2008; Weinstein, Gilmore, Szpunar, & McDermott, 2014; Yang et al., 2017), paired-associates (e.g., Weinstein, McDermott, & Szpunar, 2011; Yang et al., 2017), text passages (e.g., Wissman, Rawson, & Pyc, 2011), lecture videos (e.g., Jing, Szpunar, & Schacter, 2016; Szpunar, Khan, & Schacter, 2013), artists' painting styles (e.g., Lee & Ahn, 2018; Yang & Shanks, 2018), spatial information (Bufe & Aslan, 2018), and motor sequences (Tempel & Frings, 2019). Yue, Soderstrom, and Bjork (2015) and Yang, Chew, Sun, and Shanks (2019) demonstrated the transferability of the FTE. Yue et al. (2015, Experiment 2) for instance observed that testing on a studied lecture video on one topic (e.g., the lifecycle of a star) potentiated the learning of a new video on a completely different topic (e.g., lightning formation). Yang et al. (2019, Experiment 3) observed that testing on memory for statements about artists' contributions facilitated subsequent learning of different artists' painting styles. The FTE is observed amongst different populations, such as older adults (Pastötter & Bäuml, 2019), patients suffering from traumatic brain injury (Pastötter, Weber, & Bäuml, 2013), older elementary school children (Aslan & Bäuml, 2015), and college students with different levels of working memory capacity and test anxiety (Yang et al., in press).

Although the forward benefits of testing have been convincingly demonstrated across a variety of educational materials and in different populations, the cognitive mechanisms through which testing serves to promote new learning remain unclear (for detailed discussion, see Yang et al., 2018). Without a much deeper exploration of its underlying mechanisms, educational translation and exploitation are likely to be hindered. The main goal of the current study is to explore the cognitive underpinnings of the FTE. Below we briefly review three theories that have been put forward and then describe the rationale for the current study. Another theory, that proposes that tests boost motivation and effort, is not directly explored here but in the General Discussion we elaborate on this account and its relation to the three theories that we do assess.

Release-from-PI

Szpunar et al. (2008) postulated that the FTE results from the fact that context changes, induced by interim tests, protect new learning against interference from previously studied information – an explanation we term the *release-from-PI theory*. Interim testing on studied items updates the mental contexts in which these items are embedded, and hence these studied/tested items are associated with both a study and a retrieval context (Karpicke, Lehman, & Aue, 2014), while the subsequently studied new items are solely associated with a study context. In a subsequent test wherein participants recall the target (new) items, this context difference facilitates list segregation, enhances list item distinctiveness, delimits memory search set size, and reduces competition resulting from PI. This mechanism suggests an analogy between the effects of interpolated tests and other ways of inducing context change, such as moving to a new learning environment or interpolating a time interval (Jang & Huber, 2008; S. M. Smith & Vela, 2001).

Many studies offer support for the release-from-PI explanation by showing that the FTE tends to be correlated with reduced prior list intrusions (e.g., Aslan & Bäuml, 2015; Bufe & Aslan, 2018; Nunes & Weinstein, 2012; Pastötter & Bäuml, 2014; Pastötter et al., 2013; Pierce, Gallo, & McCain, 2017; Szpunar et al., 2008; Weinstein et al., 2014; Weinstein et al., 2011; Yang et al., 2019; Yang et al., 2017). Specifically, these studies consistently observed that interim testing concurrently potentiated recall of new information and substantially reduced the number of prior list intrusions (see the above discussion of Szpunar et al., 2008, for an illustration). A seductive inference from the simultaneous occurrence of reduced prior list intrusions and enhanced recall of new information is a causal connection, namely that interim testing enhances new learning by protecting it against PI.

Without conducting analyses to test specifically whether prior list intrusions mediate the enhancing effect of interim testing, it is, however, premature to conclude that release-from-PI is truly the mechanism underlying the FTE. As noted by many researchers (e.g., Dunlosky & Mueller, 2016; Gunzler, Chen, Wu, & Zhang, 2013; Montoya & Hayes, 2017; Pieters, 2017), mediation analysis is an essential technique to identify whether an independent variable (e.g., interim testing) exerts its effect on a dependent variable (e.g., retrieval of new information) via a mediator (e.g., release-from-PI). To our knowledge, no research has conducted a mediation analysis to examine whether release-from-PI is a potential source of the FTE (and if it is, to what extent the FTE should be attributed to release-from-PI). A possible reason for this lacuna is that the sample sizes in previous FTE studies were relatively small (e.g., 12 participants in each group in Szpunar et al., 2008). It is well-established that mediation analysis requires large sample sizes to achieve acceptable levels of statistical power (Figgou & Pavlopoulos, 2015; Schoemann, Boulton, & Short, 2017). Going beyond prior studies, the current study recruited a large sample (over 1,000 participants) to provide sufficient statistical power to determine whether release-from-PI contributes to the FTE.

Strategy-change

Strategy-change theory hypothesizes that prior tests on studied information teach learners how to study, and they then develop and adopt more effective study and retrieval strategies during subsequent learning and test phases (Cho, Neely, Crocco, & Vitrano, 2016; Cho & Powers, 2019; Soderstrom & Bjork, 2014). A study by Chan, Manley, et al. (2018) illustrates some of the evidence supporting this hypothesis. Chan and colleagues instructed a Test group and a Restudy group to study four lists of words, with each list consisting of 15 category exemplars from 5 semantic categories (i.e., with 3 exemplars from each category; e.g., *FRUIT: apple, banana, orange*; *ANIMAL: dog, cat, goat;…*), presented in a random order. The Test group was tested after studying each list, whereas the Restudy group restudied the words on each list immediately after its presentation and before moving on to the next list, and was only tested on List 4.

Chan, Manley, et al.'s List 4 test results replicated Szpunar et al.'s (2008) findings, with the Test group correctly recalling about twice as many List 4 words and committing fewer prior list intrusion errors compared with the Restudy group. More importantly, an adjusted-ratio-of-clustering analysis (ARC, varying between 0 to 1; Roenker, Thompson, & Brown, 1971), which estimates the likelihood that semantically related items follow each other during retrieval, showed superior semantic clustering in the List 4 test in the Test group (ARC = 0.58) than in the Restudy group (ARC = 0.34). This establishes that prior interim tests on Lists 1-3, by comparison with restudying, enhanced strategic processing of semantic information during List 4 encoding and retrieval.

Besides benefiting strategic processing of semantic information, testing, as opposed to restudying or no treatment, may also more effectively promote strategic processing of temporal information (Lehman & Malmberg, 2013). For instance, prior retrieval practice may teach learners to adopt more effective strategies to encode temporal information and induce superior temporal clustering during retrieval, which correspondingly boosts recall performance of new information (Sederberg, Miller, Howard, & Kahana, 2010). Strategic processing of temporal information is critical for memory formation and knowledge organization (Michon & Jackson, 1984). As an illustration, patients with frontal lobe lesions suffer from deficits in strategic processing of temporal information, leading to poor test performance in temporal order reconstruction tasks (Mangels, 1997).

Some previous studies suggest that testing does promote strategic processing of temporal information (e.g., Zacks, Hasher, Alba, Sanft, & Rose, 1984). For instance, Zacks and colleagues instructed two groups of participants to study four 36-word lists. A temporal order group took a serial order test on each list, in which words from the just-studied list were presented one-by-one and participants were instructed to recall each item's corresponding serial position (i.e., 1-36). By contrast, a free recall group took a free recall test following studying each of Lists 1-3 and then undertook a serial order test on List 4. The results showed that serial order recall accuracy increased linearly from List 1 (0.37) to List 4 (0.51) in the temporal order group, indicating that prior serial order tests enhanced temporal information processing during subsequent learning. More importantly, serial order recall accuracy in the free recall group’s List 4 order test (0.47) was better than that in the temporal order group’s List 1 order test (0.37), suggesting that free recall tests, similar to serial order tests, can promote subsequent processing of temporal information.

Semantic clustering is not a relevant factor when semantically-unrelated materials are used, as in the experiment we report below, but the above discussion points to the inference that strategic processing of temporal information may be important for knowledge organization and retention and may benefit from testing, in turn contributing importantly to the FTE. To test this hypothesis, the current study explores whether testing enhances temporal processing by comparing the level of temporal clustering (i.e., the extent to which temporally proximal items follow each other during retrieval) in the target list recall between a Test and a Restudy group. If the answer is affirmative, a mediation analysis will be conducted to investigate whether and to what extent the difference in temporal clustering mediates the FTE.

In summary, although the findings of Chan, Manley, et al. (2018) may reflect a role for strategy-change in the FTE, their findings are specific to semantic processing. Moreover, because Chan et al. (2018) did not conduct mediation analyses, it is unwarranted to draw any firm conclusions about whether strategy-change is likely to play a role in the FTE. The current study aims to fill this gap by investigating via mediation analysis whether interim testing boosts new learning through enhancing strategic processing of temporal information.

Reset-of-Encoding

As discussed above, the release-from-PI theory assumes that the FTE principally results from the influence of prior interim tests on subsequent retrieval of new information (that is, prior interim tests protect retrieval of new information from PI). By contrast, another recently proposed theory – the reset-of-encoding theory – focuses on the influence of prior interim tests on subsequent encoding of new information (Pastötter, Engel, & Frings, 2018). Specifically, the reset-of-encoding theory proposes that memory load gradually increases across the course of a study session in the absence of interim testing. Encountering a test following the study of each list reduces memory load, which in turn “resets” the subsequent encoding process and provides greater capacity for encoding and storage of new information.

A straightforward prediction of this theory is that a large enhancement should occur at the early phase of new encoding (i.e., a large enhancing effect on primacy items), but the enhancement ought to decrease at later phases of new encoding (i.e., a smaller or even null effect on non-primacy items). The rationale for this prediction is that encoding reset, induced by prior tests, makes new learning as effective as prior learning; as new learning takes place, memory load gradually increases, attenuating the encoding reset benefits and leading to a smaller and smaller enhancement for subsequently studied items.

Supporting evidence for this mechanism comes from a study by Pastötter et al. (2018), in which participants were instructed to study three 12-word lists in two conditions: Test *vs*. Restudy. In the Test condition, participants took a free recall test after studying each list; by contrast, participants restudied Lists 1 and 2 but were tested on List 3 in the Restudy condition. Again, Pastötter et al. replicated the FTE, with superior List 3 recall in the Test condition than in the Restudy condition. To test the reset-of-encoding theory, Pastötter et al. conducted a serial position analysis across the List 3 items, which showed that serial position significantly modulated the FTE, with a larger recall enhancement for the early List 3 items (i.e., Items 1-4) compared to the middle (i.e., Items 5-8) and end (i.e., Items 9-12) items.

Although Pastötter et al.’s serial position results suggest a selective enhancement effect of interpolated testing on new learning, they cannot be taken as direct evidence supporting the reset-of-encoding theory for at least two reasons. The first is that the selective enhancement effect on new learning can be readily accounted for by another explantion – *output order* (Dalezman, 1976; Yang et al., in press). Yang et al. (in press) recently reported that interpolated testing significantly affects output order. In their study, Yang et al. instucted a Test and a Restudy group to study five 18-word lists, with the Test group tested on each list and the Restudy group restudying Lists 1-4 and undertaking a free recall test on List 5.

Following the List 5 test, both groups took a final cumulative test in which they recalled as many words as they could from all five lists in any order they liked. The results showed that the Test group preferred to initiate cumulative test recall with List 1 words (i.e., they were more likely to organize recall in the order of encoding), whereas the Restudy group preferentially recalled List 4 words first (i.e., they were more likely to organize recall in the reverse order of encoding). Overall, Yang et al.'s findings suggest that retrieval practice alters output order by making learners more likely to organize recall in the order that list items are encoded (an effect consistent with strategy-change theory, of course).

Output order can explain Pastötter et al.'s serial position results equally well as the reset-of-encoding account. For instance, according to Yang et al. (in press), Pastötter et al.'s Test group might have initiated List 3 recall with primacy items (i.e., Items 1-4) whereas the Restudy group might have primarily recalled the recency items first (i.e., Items 9-12). (To foreshadow, this assumption is corroborated in the current study.) According to the *output interference effect* (that is, first-recalled items impair subsequent recall of others; A. D. Smith, D'Agostino, & Reid, 1970), recalling early items (i.e., Items 1-4) first would have hindered recall of not-yet-recalled ones (i.e., Items 5-12) in the Test group, and recalling the recency items (Items 9-12) first would have interfered with recall of Items 1-8 in the Restudy group. These output interference consequences could jointly have led to the superior recall of the early List 3 items in the Test group, as observed by Pastötter et al. (2018). Hence, Pastötter et al.'s serial position results do not straightforwardly support the reset-of-encoding theory.[1]

The second reason is that Pastötter et al. (2018) did not conduct mediation analysis to directly assess the role of reset-of-encoding in the FTE. Concurrently observing a selective enhancing effect of testing on new learning for early items and an FTE does not establish that reset-of-encoding contributes to the FTE (Dunlosky & Mueller, 2016). Hence, further mediation tests are required to justify or disprove the potential contribution of reset-of-encoding. Going beyond Pastötter et al. (2018), the current study

[1] At face value, this output order theory also predicts superior recall of recency items in the control group, which Pastötter et al. (2018) did not observe. However other mechanisms (such as test-enhanced motivation or release-from-PI) may cancel out any such effect.

aims to test the reset-of-encoding theory in a more direct way by exploring the modulating role of serial position in the FTE with output order as a controlled variable. Specifically, we test the reset-of-encoding theory by investigating whether the selective enhancing effect of interim testing persists when output order is controlled. More importantly, the current study develops a variety of indices to represent reset-of-encoding and subjects them to mediation analyses (see below for details).

Summary of Hypotheses

Current evidence for the release-from-PI theory is inconclusive because no studies have conducted mediation analyses to determine whether (and if so, to what extent) release-from-PI (indexed by prior list intrusions) plays a role in the FTE. Chan, Manley, et al. (2018) provided suggestive evidence supporting the strategy-change theory by showing that interim testing potentiates semantic clustering during subsequent learning, but they did not conduct mediation analyses and no research has investigated the potential role of temporal processing strategy-change in the FTE. Pastötter et al.'s (2018) serial position results do not constitue clear-cut support for the reset-of-encoding explanation because of (1) potential confounding with output order, and (2) lack of mediation tests. Furthermore, it is unknown which mechanism plays a more important role in the FTE. In sum, our theoretical understanding of the FTE is still in its infancy, and the current study aims to further test the proposed theories by conducting a large sample experiment and employing more advanced analytic methods.

Method

Participants

To test the roles of release-from-PI, temporal processing strategy-change, and reset-of-encoding mechanisms in the FTE, we pre-planned to recruit over 1,000 participants in a large sample study, which is a requirement for mediation analysis.[2] Accordingly, 1,075 participants were recruited from Fuqing

[2] Another reason for recruiting a large number of participants was that this project also aimed to investigate individual differences in the FTE. Because previous studies observed inconsistent findings about individual

Branch of Fujian Normal University. Data from 42 individuals were not recorded due to computer errors, and we also excluded data from one participant who made notes during the task, leaving a final sample of 1,032 participants. Note that other aspects of the data from this project, regarding individual differences in the FTE, are reported in Yang et al. (in press).

Participants were randomly assigned to a Test (518 participants) and a Restudy (514 participants) group. Their mean age was 18.63 years (*SD* = 1.10; 96 did not report their age). Six hundred and fifty-nine were female, 284 were male, and the remaining 89 did not report their gender. All were native Chinese speakers. They were tested either individually or in groups of up to 20 in a quiet laboratory room. They participated either for course credit, for monetary compensation, or voluntarily. The Ethics Committee at School of Education, Fuqing Branch of Fujian Normal University, approved this study.

Materials

For the multi-list learning task, 90 two-character high-frequency and semantically unrelated Chinese words were selected from Liu and Reichle (2017; available at https://osf.io/fp3yw/). Word frequency ranged from 51.98 to 768.09 per million (*M* = 132.47; *SD* = 111.00), and the number of strokes ranged from 10 to 21 (*M* = 14.88; *SD* = 2.68). To prevent any item selection effects, for each participant the computer randomly assigned the words into Lists 1-5 and presented them in a random order.

Procedure

The study consisted of three tasks: questionnaires, a multi-list learning task, and a working memory task. Specifically, before commencing the multi-list learning task, participants completed a set of questionnaires to measure various psychological characteristics (e.g., trait test anxiety, attitude to failure, mindset of intelligence, and so on), and after completing the multi-list task, they undertook an Operation Span (OSPAN) task to measure their working memory capacity (Unsworth, Heitz, Schrock, & Engle,

differences in test-enhanced learning, we decided to employ a large sample size (over 1000 participants) to obtain more robust findings (for detailed discussion, see Yang et al., in press).

2005). Because the data collected by the questionnaires and the OSPAN task are not relevant to the current research questions, we do not discuss them further.[3]

In the multi-list learning task, participants were informed that they would study five lists of words in preparation for a final cumulative test, during which they would be asked to recall as many words as they could from all five lists. They were warned at the outset that, after studying each individual list, the computer would randomly decide either to give them a memory test or to offer them a restudy opportunity before moving on to the next list. In fact, the test decisions were predetermined, with the Test group tested on each of Lists 1-5 and the Restudy group restudying Lists 1-4 prior to being tested on List 5.

In the List 1 study phase, 18 words were presented one-by-one, for 2 sec each, in a random order. A cross sign was presented for 0.5 sec between the presentation of two words to mark the interstimulus interval (ISI). After studying List 1, both groups solved as many simple math problems (e.g., *23 + 36 = ?*) as they could for 1 min. Next, the Test group were instructed to recall as many words as they could from the just-studied list (List 1) in 1 min.[4] By contrast, all words reappeared one-by-one, for 2 sec each, and in a new random order, for the Restudy group to restudy. The procedures for Lists 2-4 were the same as for List 1, except that participants studied new words in each list.

After the completion of List 4, both groups studied List 5 and solved math problems for 1 min. Then both groups were informed that the computer had decided to test them on List 5 (i.e., they would be required to recall as many List 5 words as they could), and they rated how anxious they were regarding the upcoming test. Then both groups took the List 5 interim test, during which they had unlimited time to

[3] Yang et al. (in press) report analyses which relate the multi-list learning data to these individual difference measures.

[4] We limited the duration (i.e., 1 min) of the List 1-4 interim tests in order to roughly equate the task (Test/Restudy) duration between groups. The Restudy group spent 45 s restudying words and hence it would have been possible to allocate the same amount of time on interim tests to completely equate the task duration between groups. However, because we were concerned that participants might be unable to complete their recall in 45 s, we extended recall time to 1 min, following previous studies (e.g., Szpunar et al., 2008).

recall as many List 5 words as they could.[5] In summary, both groups studied List 5 words once, engaged in a 1-min distractor task, reported their test anxiety about the upcoming test, and finally were tested on this list. The List 5 interim test was the target test to measure the FTE.

Following the List 5 interim test, both groups reported how anxious they were about the final cumulative test[6] and then completed it, during which they recalled as many words as they could from all five lists in any order. The cumulative test was self-paced. No feedback was provided in the interim or cumulative tests.

Results and discussion

List 5 interim test recall was at floor for many participants in the Restudy group, leading to missing data for several measures (see below for details). In addition to the main analyses reported below, we employed a variety of supplemental methods for handling missing data to address this issue. These are referred to in the main text but, for the sake of conciseness, are descibed in detail in the Appendix. Note that, regardless of how missing data were treated, all results show the same patterns.

In addition to standard significance tests, we conducted Bayesian analyses via JASP (JASP Team, 2020).

Interim test recall

Interim test recall for each of Lists 1-5 is reported in Table 1. For the Test group, a Bayesian repeated measures analysis of variance (ANOVA) found no significant fluctuation of interim test recall

[5] We did not control the duration of the List 5 interim test due to the concern that the Restudy group, having no experience with the interim tests, might not be able to complete their recall within 1 min. Indeed, the Restudy group spent more time (M = 128.51 s, SD = 80.90) on the List 5 interim test than the Test group (M = 98.56 s, SD = 44.23), difference = 29.95 [22.00, 37.91], $t(1030) = 7.39$, $p < .001$, Cohen's d = 0.460, BF_{10} = 1.8e+10. Another noteworthy point is that, even though the Test group spent longer than 1 min on the List 5 interim test, this longer time did not significantly enhance recall compared to recall on the List 1-4 interim tests (see below for details). The reason might be that, as documented in numerous studies, participants were able to retrieve most of their remembered items at the beginning of the free recall test, with correct recall rapidly declining across the later part of the test (see, Bäuml & Kliegl, 2013, Figure 1 for an illustration of the exponentially decreasing relationship between correct recall and recall latency).

[6] Yang et al. (in press) report the test anxiety results and their relationship to the FTE.

across lists, $F(4, 2{,}068) = 0.68$, $p = .606$, $\eta_p^2 = .001$, $BF_{10} = 6.0\text{e-}4$. As shown in Figure 1A, the Test group correctly recalled more words than the Restudy group in the List 5 interim test, difference = 4.51 [4.03, 5.00], $t(1030) = 18.21$, $p < .001$, Cohen's $d = 1.13$, $BF_{10} = 5.6\text{e+}60$, revealing a highly robust FTE with a very large effect size.

Prior list intrusions

Prior list instrusions across lists are reported in Table 1. In the Test group, prior list intrustions linearly increased from the List 2 to the List 5 interim test, $F(3, 1{,}551) = 35.72$, $p < .001$, $\eta_p^2 = .065$, $BF_{10} = 5.3\text{e+}19$, indicating that interim testing does not completely prevent the build-up of PI across a study session. The Test group experienced far fewer prior list intrusions, however, than the Restudy group in the List 5 interim test, difference = -5.23 [-4.76, -5.69], $t(1030) = -22.10$, $p < .001$, $d = -1.38$, $BF_{10} = 1.2\text{e+}85$ (see Figure 1B), revealing extremely strong evidence that interim testing reduces the build-up of PI across lists.

Cumulative test recall

In the cumulative test, the Test group ($M = 18.97$; $SD = 10.32$) correctly recalled more List 1-5 words than the Restudy group ($M = 13.75$; $SD = 11.60$), difference = 5.22 [3.88, 6.57], $t(1030) = 7.65$, $p < .001$, $d = 0.48$, $BF_{10} = 1.1\text{e+}18$ (see Figure 1C). Given that the current study is focused on the FTE, we do not discuss the cumulative test results further. Interested readers can consult Yang et al. (in press).

Release-from-PI

To test the relationship between correct recall and PI (indexed by prior list intrusions) in the List 5 interim test, a linear regression analysis was conducted for each group, in which List 5 interim test recall was regressed on prior list instrusions across participants. There was a negative relationship between prior list intrusions and correct recall across participants, slope coefficient = -0.433 [-0.488, -0.379], $p < .001$, indicating that every additional prior list intrusion reduces correct recall by 0.433 items. In addition, this positive relationship occurred in both groups (ps < .001).

To directly test the release-from-PI theory, a mediation analysis was conducted via the R *mediation* package (Imai, Keele, & Yamamoto, 2010), with List 5 interim test recall as the dependent variable, group (Test *vs*. Restudy) as the independent variable, prior list intrusions in the List 5 interim test as the mediator, and boostrap sample set to 5,000. The mediation results are reported in Table 2. Prior list intrusions accounted for 26.6% [18.9%, 35.5%], $p < .001$, of the prospective benefit of interim testing on new learning, supporting the release-from-PI theory as an account of the FTE.

Temporal processing strategy-change

Temporal clustering was quantified using the method developed by Polyn, Norman, and Kahana (2009). Temporal clustering scores (TCSs; Lohnas, Polyn, & Kahana, 2011), also known as temporal factors (Sederberg et al., 2010), were calculated for each participant based on a percentile ranking of temporal contiguity. Specifically, for each correctly recalled word (except for the last one because no recall followed it), we determined the absolute temporal distances (measured by serial position from 1-18) between the serial positions of that word and each of the not-yet-recalled ones. The TCS between the just-recalled word and the subsequently recalled one was quantified as the proportion of all other possible absolute temporal distances that were greater than the observed distance. Put differently, the TCS between the just-recalled and subsequently recalled words was computed as the proportion of all other possible temporal contiguities that were weaker than the observed one.

To illustrate, imagine that a given participant recalls Item 13, there are five words not yet recalled, and their serial positions are 6, 9, 15, 16, and 18. We first calculate the absolute differences in serial positions between Item 13 and each not-yet-recalled one to measure their temporal contiguity with Item 13. The calculated scores (i.e., subtracted from 13) for items at positions 6, 9, 15, 16, and 18 are 7, 4, 2, 3, and 5, respectively. The larger the absolute difference score in serial position, the weaker the temporal contiguity. If this participant subsequently recalls Item 15, the TCS between the just-recalled word (Item 13) and the subsequently recalled one (Item 15) is 1 because all other items’ temporal contiguities with Item 13, comprising the set {7, 4, 3, and 5}, are weaker than that of Item 15 (= 2), and

TCS is the proportion of all possible absolute distances in serial positions that are greater than the observed distance. In contrast, if the participant recalls Item 9 after Item 13, the TCS will be 0.5 as two of the remaining items (i.e., Items 6 and 18) are more distant and the other two (i.e., Items 15 and 16) are less distant than Item 9 from Item 13. In the same way, the TCS will be 0 if the subsequently recalled word is Item 6 because its temporal contiguity with Item 13 is less than (i.e., its distance is greater than) those of all other items.

The TCS for a given participant was defined as an average of the percentile ranking scores based on temporal contiguity across the correctly recalled words. Theoretically, TCSs range from 0 (the participant always transitions to the least temporally proximal item) to 1 (he/she always transitions to the most temporally proximal item). Before conducting any analyses, for each interim test and for each participant, we removed all intrusions (i.e., incorrect recall of unstudied words or prior list intrusions) and repeats (i.e., words recalled repeatedly).[7] For participants whose correct recall was 0 or 1 items, their data were excluded from these analyses because there was no successive recall and it was impossible to calculate TCSs.

The calculated TCSs for both groups are listed in Table 1. For the Test group, a Bayesian repeated measures ANOVA was conducted to explore the variation of TCSs across the List 1-5 interim tests. In total, 453 participants were included in this analysis, and the remaining 65 were excluded because they recalled fewer than 2 words in at least one of the List 1-5 interim tests, making it impossible to calculate TCSs. The ANOVA results showed that TCSs significantly varied across lists, $F(4, 1{,}808) = 18.04$, $p < .001$, $\eta_p^2 = .038$, $BF_{10} = 2.2e{+}11$. Tests of within-subjects contrasts revealed a linear increasing trend ($F(1, 452) = 39.63$, $p < .001$) as well as a quadratic trend ($F(1, 452) = 25.15$, $p < .001$). Descriptively, participants clustered their recall more in the List 2 test than in the List 1 test, and clustering thereafter remained at a roughly constant level.

[7] For repeats, the first correct recall was retained and subsequent repeats were removed. This applied to all the following analyses.

Of critical interest was the difference in List 5 TCSs. A Bayesian independent-samples *t*-test was conducted. In total, 729 participants (i.e., 485 in the Test group and 244 in the Restudy group) were included in the comparison, and the remaining 303 (i.e., 33 in the Test group and 270 in the Restudy group) were excluded because they recalled fewer than 2 words in the List 5 interim test.[8] As shown in Table 1, List 5 TCSs in the Test group were significantly larger than in the Restudy group, difference = 0.064 [0.033, 0.094], $t(727) = 4.08$, $p < .001$, $d = 0.32$, $BF_{10} = 275$, indicating that prior interim tests alter the temporal processing of items in a subsequent learning/retrieval cycle.

A regression analysis showed that List 5 TCSs positively predicted List 5 interim test recall across participants, slope coefficient = 5.725 [4.268, 7.183], $p < .001$, and this positive relation existed in both groups (*p*s < .001). These results are consistent with the findings repeatedly documented in previous studies (e.g., Lohnas et al., 2011; Sederberg et al., 2010), reflecting a positive relationship between free recall and temporal clustering and implying that temporal clustering is beneficial for information organization and retention.

To test the role of temporal processing in the FTE, a mediation analysis was conducted, identical to the previous one but with two exceptions. The first was that the mediator was replaced by TCSs. The second was that 303 participants (i.e., 33 in the Test group and 270 in the Restudy group) were excluded because they lacked TCSs. The mediation results are shown in Table 2. List 5 TCSs accounted for 12.6% [5.3%, 21.8%], $p < .001$, of the total effect of interim testing on List 5 interim test recall. This provides support for a potential role of temporal processing strategy-change in the FTE.

It should be acknowledged that the above analyses suffer from the limitation that data were excluded from 303 participants for whom List 5 TCSs were not computable. Furthermore, the proportion of missing data was greater in the Restudy than in the Test group. One approach to mitigate such data

[8] For these 729 participants, the Test group ($M = 7.633$, $SD = 3.870$) also recalled substantially more words in the List 5 interim test than the Restudy group ($M = 5.127$, $SD = 4.309$), difference = 2.51 [1.89, 3.13], $t(727) = 7.94$, $p < .001$, Cohen's $d = 0.62$, $BF_{10} = 6.9e+10$, revealing a highly robust FTE.

deletion issues is to apply a linear interpolation method to estimate non-computable TCSs (Noor, Al Bakri Abdullah, Yahaya, & Ramli, 2015). A second approach is to restrict the analysis to only a fixed percentile (specifically, the top quartile) of participants (Wilcox, 1995), as determined by their scores on the List 5 interim test. There are no missing data in the top quartile of participants. When applying these methods, all results showed the same patterns (see the Appendix for details). Overall, regardless of how the missing TCSs were treated, the results robustly and consistently support a role of temporal processing strategy-change in the FTE.

Reset-of-encoding

Figure 2A shows how the FTE (i.e., the difference in List 5 interim test recall) evolved as a function of serial position. A serial position analysis was conducted to replicate Pastötter et al.'s (2018) findings. Specifically, following Pastötter et al., we divided List 5 words into two sets, with the first third of the items (Items 1-6) comprising a primacy set and the remaining ones (Items 7-18) a non-primacy set. A Bayesian mixed ANOVA took average recall of primacy and non-primacy set items as the dependent variable, group (Test *vs*. Restudy) as the between-subjects independent variable, and serial position (primacy *vs*. non-primacy) as the within-subjects independent variable. The results revealed main effects of serial position, $F(1, 1030) = 65.55$, $p < .001$, $\eta_p^2 = .011$, $BF_{10} = 8.4e+11$, and group, $F(1, 1030) = 369.01$, $p < .001$, $\eta_p^2 = .264$, $BF_{10} = 6.3e+66$. Of critical interest, the interaction between group and serial position was also significant, $F(1, 1030) = 44.36$, $p < .001$, $\eta_p^2 = .008$, $BF_{10} = 1.4e+8$, indicating selective enhancement across serial positions. As is clearly visible in Figure 2A, against a fairly flat recall curve in the Restudy group, early List 5 items benefited more from interim testing than middle and later ones. Overall, these results replicate Pastötter et al.'s (2018) serial position findings. However, as discussed above, they cannot be taken as direct evidence for the reset-of-encoding explanation because the output order explanation might be equally valid.

To measure output order, we adopted a bidirectional pair frequency method (Anderson & Watts, 1969; Sternberg & Tulving, 1977). Specifically, we quantified output order via direction transition scores

(DTSs), which represent the extent to which a given individual organizes retrieval in a forward or backward direction (Forrin & Macleod, 2016). Imagine that a participant correctly recalls four words, and the serial positions of these words are 5, 10, 2, and 11, respectively. The direction transition from the first to the second recall was coded as +1, indicating that the transition (from position 5 to 10) was forward; the second transition (from 10 to 2) was coded as -1 to label it as backward; and the third transition (from 2 to 11) was again coded as +1 representing a forward transition. The average of these transition scores is (1 – 1 + 1)/3 = 0.333, which is then taken as the DTS for that participant. Theoretically, DTS ranges from -1 (when a given individual organizes output in a completely backward order) to +1 (when output is in a completely forward order).

Before calculating DTSs, all intrusions and repeats were removed. For a given participant and in a given interim test, the data were excluded if 0 or 1 items were correctly recalled because there was no recall direction transition. In addition, if only 2 items were correctly recalled, the data were also excluded because DTS was always -1 (completely backward) or +1 (completely forward) regardless of whether participants recalled the two words in a random or strategically controlled order.[9]

The DTSs across List 1-5 interim tests for both groups are reported in Table 1. A Bayesian repeated measures ANOVA was conducted to assess the variations in DTSs across lists in the Test group. Data from 408 participants were included in this analysis and data from the other participants were excluded because they correctly recalled fewer than 3 words in at least one of List 1-5 interim tests. The analysis found that, in the Test group, DTSs significantly varied across lists, $F(4, 1{,}628) = 8.94$, $p < .001$, $\eta_p^2 = .021$, $BF_{10} = 7{,}419$. Tests of within-subjects contrasts found a linear increasing ($F(1, 407) = 7.69$, $p = .006$) and quadratic trend ($F(1, 407) = 13.95$, $p < .001$). As with the TCSs, participants in the Test group

[9] Note that including or excluding List 5 DTSs from participants who only correctly recalled 2 items did not change the overall pattern of results. For instance, when including these participants there was stong evidence that the Test group (*M* of List 5 DTSs = 0.323, *SD* = 0.495) organized List 5 interim test recall in a more consistent forward order than the Restudy group (*M* = 0.167, *SD* = 0.729), difference = 0.16 [0.07, 0.25], $t(727) = 3.42$, $p < .001$, $d = 0.27$, $BF_{10} = 25.92$.

showed an increase in forward output organization from the List 1 test to the List 2 test, and forward organization thereafter remained at a roughly constant level.

A Bayesian independent *t*-test was conducted to compare the List 5 DTSs between groups, with 631 participants (467 in the Test group and 164 in the Restudy group) included and the remainder excluded as they recalled fewer than 3 words in List 5 interim test.[10] As shown in Table 1, List 5 DTSs in the Test group were significantly greater than in the Restudy group, difference = 0.145 [0.058, 0.232], $t(629) = 3.263$, $p = .001$, $d = 0.30$, $BF_{10} = 17.15$, supporting the assumption that participants in the Test group were more likely to organize their output in a forward order than those in the Restudy group.[11]

Supplemental analyses were conducted to compare from which position the two groups initiated their List 5 interim test recall, and the detailed results are reported in the Appendix. These corroborate the assumption that, in the List 5 interim test, participants in the Test group tended more than those in the Restudy group to begin by recalling early list items, whereas participants in the Restudy group were more likely than those in the Test group to begin by recalling the last items (for related findings, see Yang et al., in press).

Overall, the above results demonstrate that (1) the Test group organized their List 5 interim test recall in a more consistent forward order than the Restudy group, and (2) the Test group, relative to the Restudy group, preferred to initially recall early list items, whereas the Restudy group was more likely to start their recall with later list items than the Test group. Such findings are in line with the output order explanation. Hence, it is unknown whether the selective enhancement effect of interim testing across serial positions was produced by output order or reset-of-encoding.[12]

[10] For these 631 participants, the Test group ($M = 7.850$, $SD = 3.780$) also recalled more words in the List 5 interim test than in the Restudy group ($M = 6.652$, $SD = 4.532$), difference = 1.20 [0.49, 1.91], $t(629) = 3.31$, $p < .001$, Cohen’s $d = 0.30$, $BF_{10} = 19.77$, revealing a highly robust FTE.

[11] A regression analysis showed that List 5 DTSs positively predicted List 5 interim test recall, slope coefficient = 1.667 [1.041, 2.293], $p < .001$, confirming that the more participants organized recall in a forward order, the more words they recalled in the List 5 interim test.

[12] It is worth noting that reset-of-encoding and output order are not mutually exclusive. In addition, it should be acknowledged that the fact that the Test group was more inclined to initiate List 5 interim test recall with the

Figure 2B shows the serial position curves for each group in this subset of participants for whom DTSs could be calculated. To directly test whether reset-of-encoding contributes to the FTE, we conducted a Bayesian mixed ANOVA with average recall of primacy and non-primacy items as the dependent variable, serial position (primacy *vs*. non-primacy) as the within-subjects independent variable, group as the between-subjects variable, and DTSs as a covariate. This found main effects of group ($F(1, 628) = 13.21$, $p < .001$, $\eta_p^2 = .020$, $BF_{10} = 60.91$), serial position ($F(1, 628) = 11.07$, $p < .001$, $\eta_p^2 = .005$, $BF_{10} = 4.4e+10$), and DTSs ($F(1, 628) = 28.29$, $p < .001$, $\eta_p^2 = .042$, $BF_{10} = 73{,}270$). Importantly, the interaction between study strategy and serial position remained significant when DTSs were controlled, $F(1, 628) = 11.67$, $p < .001$, $\eta_p^2 = .008$, $BF_{10} = 1{,}098$, indicating that interpolated testing tends to selectively enhance recall of new items over and above confounding with output order. (Readers should be cautious about these results for reasons that will be elaborated in the General Discussion.)

It should be noted that including List 5 DTSs as a covariate significantly enhanced goodness of model fit ($BF_{10} = 73{,}270$), which also reduced the explanatory power (i.e., effect size) of the interaction between group and serial position ($\eta_p^2 = .021$ *vs*. $\eta_p^2 = .008$). Hence the selective enhancement of interim testing on new learning should at least partially be attributed to changes in output order induced by interim testing, and reset-of-encoding is not the only explanation.

Again, the above analyses suffered from data deletion problems, and we therefore used the linear interpolation and percentile methods to mitigate these issues. As shown in the Appendix, all results showed the same patterns when the missing DTSs were imputed by linear interpolation and when the analysis was restricted to the top quartile of participants.

primacy items might result from the fact these items were encoded more strongly as a result of encoding reset. On the other hand, the selective enhancement effect of interim testing, which is taken to support the reset-of-encoding explanation, might completely result from output sequence change (rather than any reset-of-encoding). The current study was not primarily designed to disentangle the influences of reset-of-encoding and output order, and future research on this issue is required.

Even though the above findings demonstrated that the selective enhancing effect of interim testing on new learning survived when output order was controlled, it is still necessary to subject reset-of-encoding to mediation tests to determine its role in the FTE. We note that we developed a variety of indices to represent reset-of-encoding, but all of them failed to provide evidence supporting the reset-of-encoding view. For the sake of conciseness, below we report the results from one measure, logistic regression coefficients. The results from four other indicies are reported in the Appendix.

As shown in Figure 2A, the decrease of test-potentiated new learning mainly occurred across Items 1-8. We hence took recall of the first 8 items and computed a logistic (0 = unrecalled; 1 = recalled) regression across serial positions for each participant, and took the regression slope coefficient as an idex of reset-of-encoding. The logic is that the more negative the slope coefficient, the greater the reset-of-encoding benefit for primacy items. A Bayesian independent *t*-test showed that the slope coeffcients were significantly more negative in the Test (*M* = -3.001; *SD* = 16.185) than in the Restudy (*M* = -0.884; *SD* = 16.700) group, difference = -2.117 [-4.125, -0.108], *t*(1030) = 2.067, *p* = .039, *d* = -0.129, BF_{10} = 0.571, even though the Baysesian evidence somewhat supported the null hypothesis. A mediation analysis showed that although these slope coefficients signficantly mediated the FTE, proportion explained = -1.4% [-3.4%, -0.04%], *p* = .034 (see Table 2), but the mediating effect was in the exactly reverse pattern as the reset-of-encoding theory predicts (i.e., when controlling the mediating effect of reset-of-encoding, the FTE became larger rather than smaller). Another mediation analysis, which only included the 631 participants for whom List 5 DTSs were computable and took their List 5 DTSs as a controlled variable, again showed that the mediation effect was in the reverse direction to the reset-of-encoding theory's prediction, proportion explained = -19.1% [-94.0%, -4.0%], *p* =.013.

Release-from-PI and temporal processing strategy-change

In the above analyses, we found that both release-from-PI and temporal processing strategy-change contributed to the observed FTE, whereas the reset-of-encoding view received less support. Next, we conducted a further mediation analysis with List 5 interim test recall as the dependent variable, group as the independent variable, and prior list intrusions in the List 5 interim test and List 5 TCSs as the mediators. This mediation analysis was conducted via the SPSS *PROCESS* (Version 3.2) package, with boostrap sample set to 5,000 (Montoya & Hayes, 2017).[13]

This mediation analysis was performed with three aims. The first was to determine which mechanism (release-from-PI *vs*. temporal processing strategy-change) account for a larger proportion of the FTE. The second was to quantify to what extent these two mechanisms can jointly account for the FTE. Lastly, but importantly, it was conducted to determine whether one mechanism can independently mediate the FTE when the other is controlled. These two machanisms' contributions to the FTE may be mutually dependent. For instance, interpolated testing protects new learning from PI, which enables individuals to more effectively process temporal information during new learning. In addition, fewer prior list intrusions during retrieval may also induce superior temporal organization of the target items. Of course, it is also possible that fewer prior list intrusions during retrieval in the Test group might result from temporal processing strategy-change. For instance, better temporal organization of new items can more effectively protect them from prior list intrusions. Indeed, a correlation analysis showed that prior list intrusions in the List 5 interim test were negatively related to List 5 TCSs, r = -0.201, p < .001.

In total, 729 participants (485 in the Test group and 244 in the Restudy group) were included in the following mediation analysis. The remaining 303 (33 in the Test group and 270 in the Restudy group) lacked TCSs (i.e., correct recall in the List 5 interim test < 3) and were excluded. The detailed results are reported in Table 2. The indirect effect through release-from-PI (indexed by prior list intrusions) was 1.990 [1.549, 2.488], confirming a mediating role of release-from-PI in the FTE. The indirect effect via

[13] Another mediation analysis was conducted in which three mediators were included: prior list intrusions in the List 5 interim test, List 5 TCSs, and the difference in recall between primacy and non-primacy items (an index of reset-of-encoding). The results were largely the same as those reported above. There was a significant indirect effect through release-from-PI (1.972 [1.529, 2.449]), a significant indirect effect through temporal processing strategy-change (0.262 [0.112, 0.448]), and a significant difference between these two indirect effects, difference = 1.711 [1.232, 2.217]. Importantly, there was a significantly negative indirect effect through release-from-PI (-0.090 [-0.198, -0.009]), again not supporting the main proposal of the reset-of-encoding theory. In addition, this indirect effect was significantly smaller than that for release-from-PI, difference = -2.062 [-2.547, -1.622], and smaller than that for temporal processing strategy-change, difference = -0.351 [-0.562, -0.172].

temporal processing strategy-change (indexed by TCSs) was 0.261 [0.109, 0.441], confirming a mediating role of temporal processing strategy-change in the FTE. These significant findings also imply that both release-from-PI and temporal processing strategy-change independently contribute to the FTE (i.e., when one measure is controlled, the other measure still significantly mediates the FTE). The total indirect effect through release-from-PI and temporal processing strategy-change was 2.251 [1.775, 2.769], and these two mechanisms jointly accounted for about 89.8% (= 2.251 ÷ 2.506 × 100%) of the total effect of group on List 5 recall (i.e., the FTE). Furthermore, release-from-PI played a more important role than temporal processing strategy-change, difference in indirect effects = 1.729 [1.256, 2.242].

The Appendix reports results with missing List 5 TCSs imputed by linear interpolation and, separately, for only the top quartile of participants; in both cases the same patterns were obtained. In summary, regardless of how the missing TCSs were handled, the results show that release-from-PI and temporal processing strategy-change contribute independently to the FTE, and release-from-PI plays a more important role in the FTE than temporal processing strategy-change.

General Discussion

Interpolated testing of studied information strongly boosts subsequent encoding and retrieval of new information, a robust phenomenon established in dozens of prior studies (Szpunar et al., 2008; Yang et al., 2018) and replicated here. However, the cognitive underpinnings of the FTE remain elusive. For instance, although previous studies found that prior interim tests concurrently boost subsequent new learning and prevent the build-up of PI, no research has conducted mediation analyses to determine whether and to what extent release-from-PI contributes to the FTE. Although Chan, Manley, et al. (2018) showed that interpolated testing promotes semantic processing during new learning, without conducting a mediation analysis it is premature to conclude that semantic processing strategy-change is one of the sources of test-potentiated new learning. In addition, until now, no research has explored whether temporal processing strategy-change contributes to the FTE. The selective enhancement effect of interim testing on new learning across serial positions, documented by Pastötter et al. (2018), cannot be straightforwardly taken to support the reset-of-encoding account because testing may alter output order and Pastötter et al.'s study did not subject reset-of-encoding to a mediation test. Furthermore, to our knowledge, no prior research has explored which mechanism plays a more dominant role in the FTE and whether these mechanisms contribute independently to the FTE. The current study filled these gaps by testing a large sample and employing mediation methods.

Echoing a now burgeoning literature on the FTE (Chan, Meissner, et al., 2018; Pastötter & Bäuml, 2014; Yang et al., 2018), the current study demonstrated a strong FTE reflected by the finding that interim testing on Lists 1-4, by comparison with restudying, substantially boosted learning and retrieval of List 5. Moreover, the results also showed that interim testing reduces prior list intrusions in the List 5 interim test, revealing the power of interpolated retrieval practice to prevent the accumulation of PI across learning events. To test the role of release-from-PI in the FTE, a mediation analysis was conducted, which found that prior list intrusions account for about 26.6% of the observed FTE. This finding provides direct evidence supporting the release-from-PI explanation of the FTE.

It has been documented that testing induces encoding and retrieval strategy changes when the same information is restudied and re-tested (e.g., Einstein, Mullet, & Harrison, 2012; Soderstrom & Bjork, 2014; Zaromb & Roediger, 2010). The role of test-induced strategy-change in the FTE has been little explored, except in the study by Chan, Manley, et al. (2018). These researchers provided suggestive evidence that interim testing may potentiate new learning by enhancing strategic processing of semantic information. However, without support from a mediation analysis, it cannot be concluded that semantic processing strategy-change is actually one of the sources of the FTE, and moreover such clustering is only relevant to memory for semantically related materials but not to the unrelated word lists employed in the present and many other FTE experiments (Yang et al., 2017). Hence we explored whether strategic processing of temporal, rather than semantic, information is enhanced by testing and we employed mediation analysis in order to determine whether this factor is actually related to the FTE.

We hypothesized that interim testing may enhance new learning by inducing a strategic change in the way that item-to-item temporal relationships are encoded. This proposal was supported by three lines of evidence. First, interim testing on Lists 1-4 significantly enhanced temporal processing of List 5, as revealed by the significant difference in List 5 TCSs between groups. Secondly, consistent with previous studies (Lehman & Malmberg, 2013; Lehman, Smith, & Karpicke, 2014; Polyn et al., 2009; Sederberg et al., 2010), we found that temporal clustering was correlated with recall, as revealed by the positive relationship between List 5 TCSs and List 5 interim test recall. Thirdly, and most importantly, a mediation analysis showed that temporal clustering (indexed by TCSs) successfully accounted for 12.6% of the variance in the FTE, providing direct evidence for a role of temporal processing strategy-change in the FTE. Going beyound Chan, Manley, et al. (2018), we note that the current study is the first to provide mediation evidence supporting the strategy-change theory of the FTE.

Pastötter et al. (2018) proposed a reset-of-encoding explanation, which claims that testing of studied information abolishes or at least reduces memory load, resets the subsequent encoding process, and makes later learning as effective as prior encoding. The current study successfully replicated Pastötter et al.’s (2018) serial position findings, with a larger enhancement for primacy than for non-primacy items. We suspect, however, such results cannot be taken as direct support for the reset-of-encoding explanation without a careful assessment of the effects of testing on output order.

Consistent with Yang et al. (in press), the current study observed that the Test group was more likely to initiate List 5 interim test recall with early list items than the Restudy group; moreover, List 5 DTSs were significantly greater in the Test than in the Restudy group, revealing that the Test group organized List 5 interim test recall in a more consistent forward order than the Restudy group. Such findings raise the concern that the selective enhancement effect of interim testing on List 5 recall might result from an influence of testing on output order rather than via reset-of-encoding (even though these two mechanisms are not mutually exclusive and may themselves be correlated). To further test whether the selective effect at least partially results from reset-of-encoding, we ran an analysis to test whether the selective effect survived when output order (indexed by DTSs) was controlled. The answer was affirmative. [We warn readers to interprate these findings conservatively because, as discussed below, the current study only controlled output order in an indrect way.]

Even though the selective effect survived when outupt order was controlled, it must be highlighted that controlling output order significantly enhanced model goodness of fit and reduced the effect size of the interaction between group and serial order. From these findings we infer that the selective enhancing effect of interim testing on new learning is at least partially derived from changes in output order induced by interim testing, and that reset-of-encoding is not the only cause of this selective effect.

Although the documented findings imply that the selective enhancing effect might partially result from reset-of-encoding, it says little about whether reset-of-encoding actually explains the observed FTE. Hence, a variety of indices were constructed to represent the magnititude of reset-of-encoding and several analyses were performed to determine their mediating effects, but none of them provided evidence supporting the reset-of-encoding theory. Even though the documented results provide little support for the reset-of-encoding theory, it would be premature to conclude that this mechanism plays no role in the FTE, because the current study only controlled output order in an indirect way (see below for detailed discussion). Future research could usefully control output order in a more direct way to further evaluate the validity of this theory.

After establishing the contributions of release-from-PI and temporal processing strategy-change, we conducted a final mediation analysis with two mediators (i.e., prior list intrusions and TCSs in the List 5 interim test) to explore (1) which mechanism plays a more important role in the FTE, (2) to what extent these two machanisms jointly account for the FTE, and (3) whether these two mechanisms independently contribute to the FTE. The results showed that release-from-PI accounted more strongly for the FTE than temporal processing strategy-change. This interesting result should of course not be over-generalized. For the present task (learning lists of Chinese words), it appears that release-from-PI is somewhat more important than strategy-change, but this may be a consequence of specific features of word learning with these materials. Whether similar patterns would occur for other forms of learning (e.g., paired-associates) or with other materials (e.g., face-name pairs) must await future research. The results also revealed that release-from-PI and temporal processing strategy-change jointly accounted for a major proportion (about 89.9%) of the observed FTE after removing the data from participants for whom List 5 TCSs were not computable.

We have found evidence that release-from-PI and strategy-change make independent contributions to the FTE. Other mechanisms not explored here might also make unique contributions (Yang et al., 2018). The final complete theory of the effects of testing on subsequent learning may turn out to be multi-factorial. However it is natural to ask whether instead these apparently distinct mechanisms may share a deeper connection. For example, is it possible that there is some underlying process, triggered by taking a test, that simultaneously reduces PI and induces more effective temporal/semantic encoding and organization? In our view this important question motivates a call for computational modelling to be brought to bear in research on the testing effect, which has thus far been very limited. We offer one outline speculation about how such an agenda could be taken forward. Karpicke et al. (2014) have offered an episodic context theory of testing effects in which context reinstatement, context updating, and restriction of the search set are all facilitated by retrieval practice. These mechanisms could prove to be adequate to explain all three of the main findings noted above. For example, greater differentiation between list contexts, induced by taking an interim test, might explain the reduction of PI observed in experiments on the FTE. Instantiating these or other mechanisms in a formal model is likely to considerably assist researchers in understanding the forward benefits of testing.

Another observation is that although the present work compared a range of prominent theories, there are others which it does not consider (see Yang et al., 2018; for a review of 8 theories that have been proposed). Prominent amongst them are motivation/effort accounts of the FTE (Cho et al., 2016; Weinstein et al., 2014; Yang et al., 2018). These theories hypothesize that prior interpolated tests provide experience of recall failure, induce dissatisfaction about poor test performance, and raise test expectancy (i.e., expecting to be tested on the to-be-studied materials), which jointly motivate learners to commit more effort to encode and retrieve new information. Although the current study did not measure participants' study motivation/effort, motivational/effort theories can help to explain the findings. For instance, recall failures in prior tests might motivate participants to employ more effective strategies to process temporal information (i.e., temporal processing strategy-change) and exert greater effort toward encoding new items, which in turn enhanced their distinctiveness and protected them from PI (i.e., release-from-PI). Therefore, a promising direction for future research is to explore how these mechanisms interact with each other.

Although the current study focused particularly on the cognitive underpinnings of the FTE, the findings also have implications for theoretical explanations of other memory phenomena. For instance, reset-of-encoding has also been proposed to account for the enhancement effect of forgetting on new learning (Pastötter, Kliegl, & Bäuml, 2012, 2016). Specifically, in the list-method directed forgetting procedure, participants first study List 1 and are instructed to either remember or forget it, after which they study List 2. In a subsequent recall test, the forgetting instruction leads to inferior recall of List 1 together with superior recall of List 2. Importantly, the enhancement effect of forgetting on learning of List 2 is selective, with larger enhancement observed for primacy items than for non-primacy ones (Pastötter et al., 2012, 2016). Pastötter et al. proposed reset-of-encoding to account for this selective enhancement effect: The forgetting instruction resets the subsequent encoding of List 2, leading to a larger enhancement for early List 2 items (for a review, see Pastötter, Tempel, & Bäuml, 2017).

The current study observed that interpolated testing alters output order, which may contribute to the selective enhancement effect of interim testing on new learning. It is possible that a forgetting instruction may also affect output order, which in turn (partially) contributes to the selective enhancement effect of forgetting on new learning. Therefore, future research is needed to reassess the role of reset-of-encoding in the selective enhancement effect of forgetting on new learning with output order controlled.

More importantly, future research should aim to develop other indices to test the mediating role of reset-of-encoding.

Limitations

Although the current study recruited a large sample to test the mechanisms underlying the FTE and most of the results are strong (as revealed by the BF_{10}s > 10 and *p*s < .001), three limitations must be acknowledged. First, all participants in the current study were native Chinese speakers and all stimuli were Chinese words. Future research could profitably test the generalization of the findings in different countries and languages.

Second, taking DTSs as a control variable to explore the selective enhancement effect of interim testing on new learning across serial positions is an indirect method to assess the selective enhancing effect of reset-of-encoding. Future studies should control output sequence in a more direct way. For instance, after participants study each list of words (e.g., *apple, computer, hand* …), a cued recall test (e.g., *a*_____, *c*_____, *h*____ …) could be administered to control output order (Roediger, 1974; A. D. Smith, 1977). The cued recall test will partial out the influence of output interference and provide a direct path to measure any selective enhancing effect induced by reset-of-encoding.

Third, for many participants, their List 5 interim test recall was at floor (i.e., correct recall < 3), leading to List 5 TCSs and TDSs not being computable. In addition, missing List 5 TCSs and DTSs occurred more frequently in the Restudy group than in the Test group, leading to non-equivalent numbers of participants excluded from data analyses. Even though we imputed missing data via linear interpolation and conducted analyses restricted to the top quartile of participants, and the re-calculated results showed the same patterns (see the Appendix), it must be acknowledged that these methods are imperfect (Noor et al., 2015). Future research in which the floor effect in List 5 interim test recall is avoided (or at least alleviated) – for instance, by employing fewer (e.g., 3 rather than 5) word lists – would be useful (Pastötter & Frings, 2019).

Concluding Remarks

How to sustain learning efficiency across a prolonged period of study is a challenge for all learners. The current experiment, along with many other FTE studies, demonstrated that inserting interim tests into a learning session can maintain learning efficacy, suggesting that learners and instructors should endeavor to interperse retrieval practice into a prolonged learning episode whenever possible. More importantly, the current study demonstrated that release from proactive interference and strategic changes in the encoding of temporal information jointly contribute to test-potentiated new learning, whereas the reset-of-encoding account receives less support.

References

Abbott, E. E. (1909). On the analysis of the factor of recall in the learning process. *The Psychological Review: Monograph Supplements, 11*, 159-177. doi:http://dx.doi.org/10.1037/h0093018

Anderson, R. C., & Watts, G. H. (1969). Bidirectional associations in multi-trial free recall. *Psychonomic Science, 15*, 288-289. doi:https://doi.org/10.3758/BF03336303

Aslan, A., & Bäuml, K. H. T. (2015). Testing enhances subsequent learning in older but not in younger elementary school children. *Developmental Science,19*, 992-998. doi:https://doi.org/10.1111/desc.12340

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

Bufe, J., & Aslan, A. (2018). Desirable difficulties in spatial learning: Testing enhances subsequent learning of spatial information. *Frontiers in Psychology, 9*, 1701-1701. doi:10.3389/fpsyg.2018.01701

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

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

Cho, K. W., Neely, J. H., Crocco, S., & Vitrano, D. (2016). Testing enhances both encoding and retrieval for both tested and untested items. *Quarterly Journal of Experimental Psychology*, 1-60. doi:10.1080/17470218.2016.1175485

Cho, K. W., & Powers, A. (2019). Testing enhances both memorization and conceptual learning of categorical materials. *Journal of Applied Research in Memory and Cognition, 8*, 166-177. doi:https://doi.org/10.1016/j.jarmac.2019.01.003

Dalezman, J. J. (1976). Effects of output order on immediate, delayed, and final recall performance. *Journal of Experimental Psychology: Human Learning and Memory, 2*, 597-608. doi:10.1037/0278-7393.2.5.597

Dunlosky, J., & Mueller, M. (2016). Recommendations for exploring the disfluency hypothesis for establishing whether perceptually degrading materials impacts performance. *Metacognition and Learning, 11*, 123-131. doi:10.1007/s11409-016-9155-9

Ebbinghaus, H. (1885/1913). *Memory: A contribution to experimental psychology*. New York: Teachers College, Columbia University.

Einstein, G. O., Mullet, H. G., & Harrison, T. L. (2012). The testing effect: Illustrating a fundamental concept and changing study strategies. *Teaching of Psychology, 39*, 190-193. doi:10.1177/0098628312450432

Figgou, L., & Pavlopoulos, V. (2015). Social psychology: Research methods. In J. D. Wright (Ed.), *International Encyclopedia of the Social & Behavioral Sciences (Second Edition)* (pp. 544-552). Oxford: Elsevier.

Forrin, N. D., & Macleod, C. M. (2016). Order information is used to guide recall of long lists: Further evidence for the item-order account. *Canadian Journal of Experimental Psychology, 70*, 125-138. doi:10.1037/cep0000088

Gunzler, D., Chen, T., Wu, P., & Zhang, H. (2013). Introduction to mediation analysis with structural equation modeling. *Shanghai archives of psychiatry, 25*, 390-394. doi:10.3969/j.issn.1002-0829.2013.06.009

Imai, K., Keele, L., & Yamamoto, T. (2010). Identification, inference and sensitivity analysis for causal mediation effects. *Statistical Science, 25*, 51-71. doi:10.1214/10-STS32

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. doi:10.1037/0278-7393.34.1.112

JASP Team. (2020). JASP (Version 0.14.1) [Computer software]. Retrieved from https://jasp-stats.org/

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,* 305-318. doi:10.1037/a0019902.supp

Karpicke, J. D., Lehman, M., & Aue, W. R. (2014). Retrieval-based learning: An episodic context account. *Psychology of Learning and Motivation, 61*, 237-284. doi:10.1016/b978-0-12-800283-4.00007-1

Lee, H. S., & Ahn, D. (2018). Testing prepares students to learn better: The forward effect of testing in category learning. *Journal of Educational Psychology, 102*, 203-217. doi:10.1037/edu0000211

Lehman, M., & Malmberg, K. J. (2013). A buffer model of memory encoding and temporal correlations in retrieval. *Psychological Review, 120*, 155-189. doi:10.1037/a0030851

Lehman, M., Smith, M. A., & Karpicke, J. D. (2014). Toward an episodic context account of retrieval-based learning: Dissociating retrieval practice and elaboration. *Journal of Experimental Psychology: Learning, Memory & Cognition, 40*, 1787-1794. doi:10.1037/xlm0000012

Liu, Y., & Reichle, E. D. (2017). Eye-movement evidence for object-based attention in chinese reading. *Psychological Science, 29*, 278-287. doi:10.1177/0956797617734827

Lohnas, L. J., Polyn, S. M., & Kahana, M. J. (2011). Contextual variability in free recall. *Journal of Memory and Language, 64*, 249-255. doi:10.1016/j.jml.2010.11.003

Mangels, J. A. (1997). Strategic processing and memory for temporal order in patients with frontal lobe lesions. *Neuropsychology, 11*, 207-221. doi:10.1037/0894-4105.11.2.207

Michon, J. A., & Jackson, J. L. (1984). Attentional effort and cognitive strategies in the processing of temporal information. *Annals of the New York Academy of Sciences, 423*, 298-321. doi:10.1111/j.1749-6632.1984.tb23440.x

Montoya, A. K., & Hayes, A. F. (2017). Two-condition within-participant statistical mediation analysis: A path-analytic framework. *Psychological Methods, 22,* 6-27. doi:10.1037/met0000086

Noor, N. M., Al Bakri Abdullah, M. M., Yahaya, A. S., & Ramli, N. A. (2015). Comparison of linear interpolation method and mean method to replace the missing values in environmental data set. *Materials Science Forum, 803*, 278-281. doi:10.4028/www.scientific.net/MSF.803.278

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*, 138-154. doi:10.1080/09658211.2011.648198

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

Pastötter, B., & Bäuml, K. H. (2019). Testing enhances subsequent learning in older adults. *Psychology & Aging, 34*, 242-250. doi:10.1037/pag0000307

Pastötter, B., Engel, M., & Frings, C. (2018). The forward effect of testing: Behavioral evidence for the reset-of-encoding hypothesis using serial position analysis. *Frontiers in Psychology, 9*, 1197-1197. doi:10.3389/fpsyg.2018.01197

Pastötter, B., & Frings, C. (2019). The forward testing effect is reliable and independent of learners' working memory capacity. *Journal of Cognition, 2*, 37-37. doi:10.5334/joc.82

Pastötter, B., Kliegl, O., & Bäuml, K. H. (2012). List-method directed forgetting: the forget cue improves both encoding and retrieval of postcue information. *Memory & Cognition, 40*, 861-873. doi:10.3758/s13421-012-0206-4

Pastötter, B., Kliegl, O., & Bäuml, K. H. (2016). List-method directed forgetting: Evidence for the reset-of-encoding hypothesis employing item-recognition testing. *Memory, 24*, 63-74. doi:10.1080/09658211.2014.985589

Pastötter, B., Tempel, T., & Bäuml, K. H. (2017). Long-term memory updating: The reset-of-encoding hypothesis in lst-method directed forgetting. *Frontiers in Psychology, 8*(2076). doi:10.3389/fpsyg.2017.02076

Pastötter, B., Weber, J., & Bäuml, K. H. (2013). Using testing to improve learning after severe traumatic brain injury. *Neuropsychology, 27*, 280-285. doi: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, 43*, 1063-1072. doi:10.1037/xlm0000377

Pieters, R. (2017). Meaningful mediation analysis: Plausible causal inference and informative communication. *Journal of Consumer Research, 44*, 692-716. doi:10.1093/jcr/ucx081

Polyn, S. M., Norman, K. A., & Kahana, M. J. (2009). A context maintenance and retrieval model of organizational processes in free recall. *Psychological Review, 116*, 129-156. doi:10.1037/a0014420

Roediger, H. L. (1974). Inhibiting effects of recall. *Memory & Cognition, 2*, 261-269. doi:10.3758/BF03208993

Roediger, H. L., & Karpicke, J. D. (2006). The power of testing memory: Basic research and implications for educational practice. *Perspectives on Psychological Science, 17*, 249-255. doi:https://doi.org/10.1111/j.1745-6916.2006.00012.x

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. doi:10.1037/h0031355

Rowland, C. A. (2014). The effect of testing versus restudy on retention: A meta-analytic review of the testing effect. *Psychological Bulletin, 140*, 1432-1463. doi:10.1037/a0037559

Schoemann, A. M., Boulton, A. J., & Short, S. D. (2017). Determining power and sample size for simple and complex mediation models. *Social Psychological and Personality Science, 8*, 379-386. doi:10.1177/1948550617715068

Sederberg, B. P., Miller, F. J., Howard, W. M., & Kahana, J. M. (2010). The temporal contiguity effect predicts episodic memory performance. *Memory & Cognition, 38*, 689-699. doi:10.3758/MC.38.6.689

Smith, A. D. (1977). Adult age differences in cued recall. *Developmental Psychology, 13*, 326-331. doi:10.1037/0012-1649.13.4.326

Smith, A. D., D'Agostino, P. R., & Reid, L. S. (1970). Output interference in long-term memory. *Canadian Journal of Psychology, 24*, 85-89. doi:10.1037/h0082845

Smith, S. M., & Vela, E. (2001). Environmental context-dependent memory: A review and meta-analysis. *Psychonomic Bulletin & Review, 8*, 203-220. doi:10.3758/BF03196157

Soderstrom, N. C., & Bjork, R. A. (2014). Testing facilitates the regulation of subsequent study time. *Journal of Memory and Language, 73*, 99-115. doi:10.1016/j.jml.2014.03.003

Sternberg, R. J., & Tulving, E. (1977). The measurement of subjective organization in free recall. *Psychological Bulletin, 84*), 539-556. doi:10.1037/0033-2909.84.3.539

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, 110*, 6313-6317. doi:10.1073/pnas.1221764110

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*, 1392-1399. doi:10.1037/a0013082

Tempel, T., & Frings, C. (2019). Testing enhances motor practice. *Memory & Cognition, 47*, 1270-1283. doi:10.3758/s13421-019-00932-6

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. doi:https://doi.org/10.1016/S0022-5371(74)80043-5

Unsworth, N., Heitz, R. P., Schrock, J. C., & Engle, R. W. (2005). An automated version of the operation span task. *Behavior Research Methods, 37*, 498-505. doi:10.3758/bf03192720

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*, 1039-1048. doi:10.1037/a0036164.supp

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

Wilcox, R. R. (1995). Comparing two independent groups via multiple quantiles. *Journal of the Royal Statistical Society: Series D (The Statistician), 44*, 91-99. doi:https://doi.org/10.2307/2348620

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*, 1140-1147. doi:10.3758/s13423-011-0140-7

Yang, C., Chew, S.-J., Sun, B., & Shanks, D. R. (2019). The forward effects of testing transfer to different domains of learning. *Journal of Educational Psychology, 111*, 809–826. doi:10.1037/edu0000320

Yang, C., Luo, L., Vadillo, M. A., Yu, R., & Shanks, D. R. (2020). Testing (quizzing) boosts classroom learning: A systematic and meta-analytic review. *Psychological Bulletin, Advance Online Publication*. doi:10.1037/bul0000309

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, 23*, 263-277. doi: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*, 8. doi:10.1038/s41539-018-0024-y

Yang, C., & Shanks, D. R. (2018). The forward testing effect: Interim testing enhances inductive learning. *Journal of Experimental Psychology: Learning, Memory & Cognition, 44*, 485-492. doi:10.1037/xlm0000449

Yang, C., Sun, B., Potts, R., Yu, R., Luo, L., & Shanks, D. R. (in press). Do working memory capacity and test anxiety modulate the beneficial effects of testing on new learning? *Journal of*

Experimental Psychology: Applied, Advance Online Publication. doi:https://doi.org/10.1037/xap0000278

Yue, C. L., Soderstrom, N. C., & Bjork, E. L. (2015). Partial testing can potentiate learning of tested and untested material from multimedia lessons. *Journal of Educational Psychology*, *107*, 991-1005. doi:10.1037/edu0000031

Zacks, R. T., Hasher, L., Alba, J. W., Sanft, H., & Rose, K. C. (1984). Is temporal order encoded automatically? *Memory & Cognition*, *12*, 387-394. doi:10.3758/bf03198299

Zaromb, F. M., & Roediger, H. L., 3rd. (2010). The testing effect in free recall is associated with enhanced organizational processes. *Memory & Cognition*, *38*, 995-1008. doi:10.3758/MC.38.8.995

Table 1. Descriptive results (mean and *SD*) in List 1-5 interim tests

	List 1	List 2	List 3	List 4	List 5
Interim test recall					
Test	7.11 (3.22)	7.15 (3.47)	7.23 (3.63)	7.36 (3.90)	7.17 (4.15)
Restudy	--	--	--	--	2.65 (3.80)
Prior list intrusions					
Test	--	0.19 (0.49)	0.38 (0.81)	0.41 (0.70)	0.75 (1.40)
Restudy	--	--	--	--	5.98 (5.20)
Temporal clustering scores (TCSs)					
Test	0.54 (0.15)	0.61 (0.17)	0.60 (0.16)	0.62 (0.16)	0.61 (0.16)
Restudy	--	--	--	--	0.55 (0.26)
Direction transition scores (DTSs)					
Test	0.22 (0.38)	0.37 (0.40)	0.33 (0.41)	0.36 (0.44)	0.33 (0.46)
Restudy	--	--	--	--	0.19 (0.56)

Table 2. Mediation analysis results

Mediation models	*β*	95% CI
Group – Prior list intrusions – List 5 interim test recall		
Total effect	4.514	[4.030, 4.993]
Direct effect	3.316	[2.649, 3.980]
Indirect effect through prior list intrusions	1.199	[0.902, 1.503]
Proportion explained by prior list intrusions	26.6%	[18.9%, 35.5%]
Group – List 5 TCSs – List 5 interim test recall		
Total effect	2.506	[1.854, 3.152]
Direct effect	2.190	[1.545, 2.845]
Indirect effect through List 5 TCSs	0.316	[0.134, 0.522]
Proportion explained by List 5 TCSs	12.6%	[5.3%, 21.8%]
Group – Slope coefficients across Items 1-8 – List 5 interim test recall		
Total effect	4.514	[4.032, 4.992]
Direct effect	4.576	[4.095, 5.047]
Indirect effect through slope coefficients across Items 1-8	-0.061	[-0.151, -0.002]
Proportion explained by slope coefficients across Items 1-8	-1.4%	[-3.4%, -0.04%]
Group – (Prior list intrusions, List 5 TCSs) – List 5 interim test recall		
Total effect	2.506	[1.886, 3.126]

Direct effect	0.255	[-0.468, 0.978]
Indirect effect through prior list intrusions	1.990	[1.549, 2.488]
Indirect effect through List 5 TCSs	0.261	[0.109, 0.441]
Total indirect effect	2.251	[1.775, 2.769]
Difference in indirect effects	1.729	[1.256, 2.242]

Note: Prior list intrusions refer to prior list intrusions in the List 5 interim test.

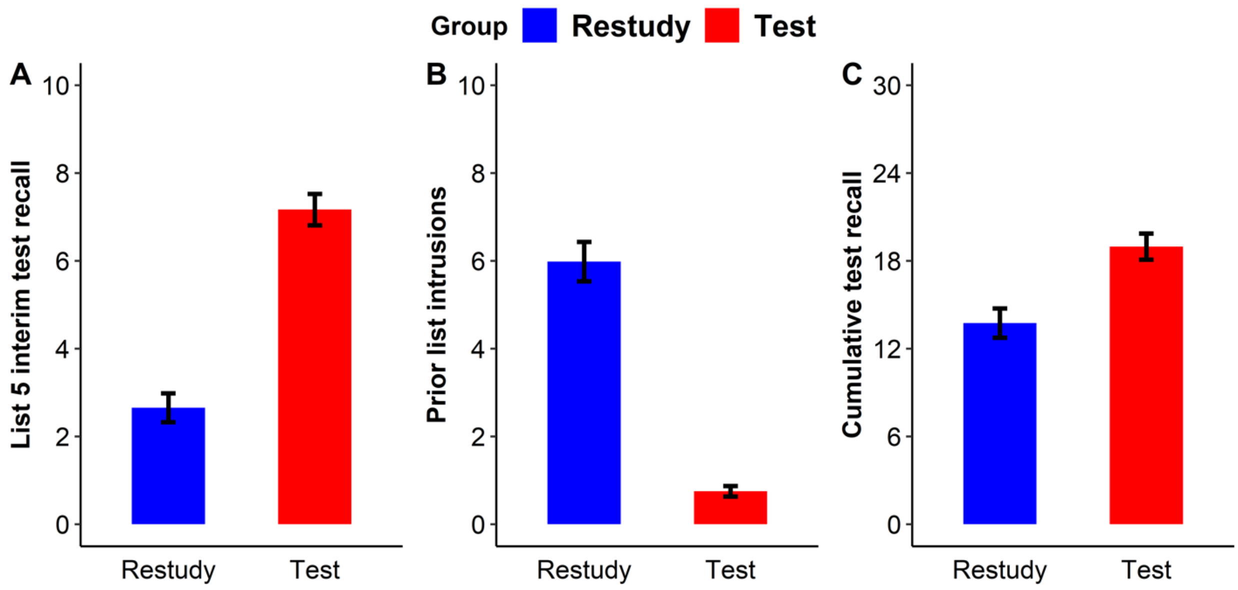

Figure 1. Panel A: List 5 interim test recall; Panel B: Prior list intrusions committed in the List 5 interim test; Panel C: Cumulative test recall. Error bars represent 95% CI.

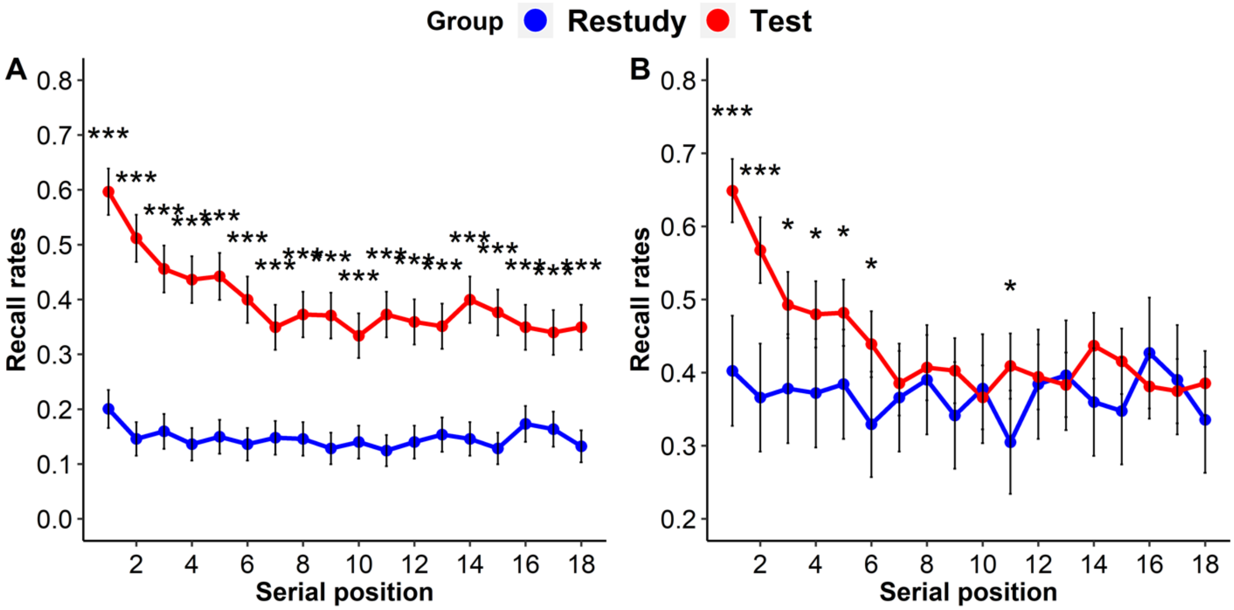

Figure 2. Panel A: List 5 interim test recall as a function of serial position in the Test and Restudy groups (all participants); Panel B: List 5 interim test recall as a function of serial position in the Test and Restudy groups (only participants who correctly recalled at least three words in the List 5 interim test). Error bars represent 95% CI. ****p* < .001; ***p* < .01; **p* < .05.

Appendix: Supplemental Analyses

Supplemental analyses to assess the role of temporal processing strategy-chage

To apply the linear interpolation method, we first regressed List 5 TCSs on List 5 interim test recall across participants, which returned a significant relationship between these two variables, slope coefficient = 0.013 [0.010, 0.017], $p < .001$, and the intercept of the regression was 0.500 [0.473, 0.526], $p < .001$. We next assigned a value of 0.500 as the List 5 TCSs for all participants who failed to recall any List 5 words, and a value of 0.513 for those who recalled only one List 5 word. We then included data from all participants to re-conduct the analyses. The results again showed that the Test group ($M = 0.604$, $SD = 0.158$) exhibited superior temporal clustering than the Restudy group ($M = 0.525$, $SD = 0.180$), difference = 0.079 [0.058, 0.100], $t(1030) = 4.08$, $p < .001$, $d = 0.32$, $BF_{10} = 3.88e+10$, and List 5 TCSs successfully accounted for 11.5% [8.3%, 15.2%], $p < .001$, of the observed FTE.

As noted in the main text, the linear interpolation method is imperfect. Hence, the following analyses were performed to further test the role of temporal processing strategy-change. Specifically, for each of the Test and Restudy group, we ranked List 5 interim test recall in an ascending order, and then only retained the top quartile of participants. If there were ties at the dividing points (i.e., the 25th percentile value), the program randomly selected which of the tied participants to include in the following analyses (Yang et al., in press). For the top quartile of participants in both groups, there were no missing List 5 TCSs. In addition, the proportions of included/excluded participants were of necessity equal between groups.

With this restriction, the Test group ($M = 12.86$, $SD = 2.89$; $N = 130$) correctly recalled more words in the List 5 interim test than the Restudy group ($M = 7.64$, $SD = 4.64$; $N = 129$), difference = 5.218 [4.273, 6.163], $t(257) = 10.87$, $p < .001$, $d = 1.35$, $BF_{10} = 4.8e+19$, reflecting a highly robust FTE. More importantly, the results again showed that the Test group ($M = 0.667$, $SD = 0.135$) exhibited superior temporal clustering than the Restudy group ($M = 0.605$, $SD = 0.198$), difference = 0.063 [0.021, 0.104], $t(257) = 2.98$, $p = .003$, $d = 0.37$, $BF_{10} = 8.72$, and List 5 TCSs successfully accounted for 8.1% [2.3%, 15.4%], $p = .008$, of the observed FTE.

Supplemental analyses to assess the role of output order

To apply the linear interpolation method, a regression analysis was conducted in which List 5 DTSs were regressed onto List 5 interim test recall across participants. The results showed that the slope coefficient was 0.025 [0.016, 0.034], $p < .001$, and the intercept was 0.106 [0.026, 0.186], $p = .010$. We hence assigned a value of 0.106 as the List 5 DTSs for all participants who failed to recall any List 5 words, a value of 0.131 for participants who recalled one List 5 word, and a value of 0.156 for participants who recalled two List 5 words.

With all 1,032 participants' data included, the results again showed that, in the List 5 interim test, the Test group ($M = 0.31$, $SD = 0.44$) organized their output order in a more consistent forward direction than the Restudy group ($M = 0.15$, $SD = 0.32$), difference = 0.17 [0.120, 0.214], $t(1030) = 6.96$, $p < .001$, $d = 0.43$, $BF_{10} = 9.96\text{e+}8$. With List 5 DTSs controlled as a covariate, the interaction between group and serial position remained significant, $F(1, 1029) = 8.48$, $p = .004$, $\eta_p^2 = .002$, $BF_{10} = 1{,}098$. Furthermore, including List 5 DTS as a covariate increased goodness of model fit ($BF_{10} = 1.71\text{e+}19$) and reduced the effect size of the interaction ($\eta_p^2 = .021$ *vs*. $\eta_p^2 = .002$).

To further mitigate missing data issues, the above analyses were reperformed, but only analyzing data from the top quartile of participants. Again, the results showed strong evidence that the Test group ($M = 0.47$, $SD = 0.42$) organized their output order in a more consistent forward direction than the Restudy group ($M = 0.21$, $SD = 0.56$), difference = 0.27 [0.144, 0386], $t(257) = 4.32$, $p < .001$, $d = 0.54$, $BF_{10} = 755$. The interaction between group and serial position in List 5 interim test recall was significant, $F(1, 257) = 4.01$, $p = .046$, $\eta_p^2 = .015$, $BF_{10} = 0.91$, even though the Baysian evidence was weak. When taking List 5 DTSs as a covariate, the interaction turned out to be non-significant, $F(1, 257) = 2.89$, $p = .091$, $\eta_p^2 = .011$, $BF_{10} = 0.80$. It is premature to draw any firm conclusion about this non-significant interaction because the sample size was relatively small (i.e., 259 participants in total) and the Bayesian evidence was weak. Crucially, including List 5 DTS as a covariate increased goodness of model fit (BF_{10} = 8.07) and reduced the effect size of the interaction ($\eta_p^2 = .015$ *vs*. $\eta_p^2 = .011$).

Supplemental analyses to measure the serial position of first correct recall

The following analyses were conducted to compare from which position the two groups initiated their recall. Specifically, we extracted each participant's first correct recall, and then determined that item's serial position (1-18). In total, 495 participants in the Test group and 356 in the Restudy group were included in this analysis; the remaining participants were excluded because they did not correctly recall any of List 5 words. Note that, for these 851 included participants, the Test group (M = 7.499, SD = 3.943) also recalled substantially more words in the List 5 interim test than the Restudy group (M = 3.829, SD = 4.048), difference = 3.67 [3.13, 4.21], $t(849) = 13.25$, $p < .001$, Cohen's d = 0.92, BF_{10} = 1.4e+33, revealing a highly robust FTE.

Figure A1 depicts the distributions of the positions of the first correct recall in the two groups. 34.9% of participants in the Test group commenced their List 5 interim test recall with the first studied word, substantially more than in the Restudy group (18.3%), proportion difference = 16.7% [10.6%, 22.7%], $\chi^2(1) = 27.814$, $p < .001$. By contrast, the Restudy group was more likely to initate their recall with the 16th (7.9%) and 17th (7.9%) words than the Test group (2.8% for the 16th word and 1.6% for the 17th word), proportion difference for the 16th word = 5.0% [1.6%, 8.4%], $\chi^2(1) = 10.149$, $p = .001$, and proportion difference for the 17th word = 6.2% [3.0%, 9.5%], $\chi^2(1) = 18.447$, $p < .001$. As shown in Figure A1, the Restudy group (7.0%) was somewhat more likely to initiate recall with the 7th word than the Test group (3.4%), difference = 3.6% [0.2%, 6.9%], $\chi^2(1) = 4.943$, $p = .026$; we assume this is sampling error.

Overall, the above results are consistent with the assumption that, in the List 5 interim test, participants in the Test group tended more than those in the Restudy group to begin by recalling early list items, whereas participants in the Restudy group were more likely than those in the Test group to begin by recalling the last items (for related findings, see Yang et al., in press).

Supplemental analyses to assess the role of reset-of-encoding

We took difference scores in average recall between Items 1-6 (primacy items) and Items 7-18 (non-primacy items) as an index of reset-of-encoding. The logic is straightforward: the larger the difference in recall between primacy and non-primacy items, the greater the benefit of reset-of-encoding for primacy items (i.e., the greater benefit primacy items receive from reset-of-encoding than non-primacy ones). Indeed, the difference scores were larger in the Test group (*M* = 0.113; *SD* = 0.290) than in the Restudy group (*M* = 0.011; *SD* = 0.193), difference = 0.102 [0.072, 0.132], *t*(1030) = 6.660, *p* < .001, *d* = 0.415, BF_{10} = 1.4e+8. However, a mediation analysis showed that these difference scores failed to significantly mediate the FTE, proportion explained = -1.0% [-3.3%, 0.9%], *p* = .337, failing to support the reset-of-encoding theory.

The above difference scores might not be sufficiently sensitive to reflect the benefit of reset-of-encoding as middle items (Items 7-12) contributed to these scores. Hence we re-calculated the difference scores in average recall between Items 1-6 and Items 13-18, which were significantly larger in the Test group (*M* = 0.113; *SD* = 0.336) than in the Restudy group (*M* = 0.005; *SD* = 0.224), difference = 0.107 [0.072, 0.142], *t*(1030) = 6.033, *p* < .001, *d* = 0.376, BF_{10} = 3.3e+6. However, a mediation analysis again showed that these re-calculated scores failed to significantly mediate the FTE, proportion explained = -0.8% [-3.0%, 1.0%], *p* = .381.

To be more cautious, we then took difference scores in average recall between Items 1-3 and Items 16-18 as an index of reset-of-encoding. Again, we found that these difference scores were larger in the Test (*M* = 0.175; *SD* = 0.444) than in the Restudy group (*M* = 0.012; *SD* = 0.290), difference = 0.163 [0.117, 0.209], *t*(1030) = 6.959, *p* < .001, *d* = 0.433, BF_{10} = 9.9e+8. But again these difference scores did not significantly mediate the FTE, proportion explained = -1.7% [-4.3%, 0.4%], *p* = .105.

Proponents of the reset-of-encoding theory may claim that the logistic regression slope coefficients, as reported in the main article, are not sufficiently reliable as each serial position reflects only one item (0 = recalled; 1 = unrecalled). Hence, the following analysis was performed. For each participant, we divided the first 8 items into 4 pairs (e.g., Pair 1 consists of Items 1 and 2), and calculated the recall average of the two items in each pair. Then we took those 4 averages and ran a linear regression across Pairs 1-4 for each participant and extracted the slope coefficient to represent the magnititude of reset-of-encoding. Again, the logic is that the more negative the slope coefficent, the larger the benefit of reset-of-encoding. A Bayesian independent *t*-test showed that the slope coefficients were more negative in the Test (*M* = -2.436; *SD* = 13.070) than in the Restudy (*M* = -0.259; *SD* = 11.829) group, difference = -2.177 [-3.700, -0.654], *t*(1030) = 2.805, *p* = .005, *d* = -0.175, BF_{10} = 3.326. However, again, a mediation analysis showed that the significant mediating effect was in the exactly reverse direction as the reset-of-encoding theory predicts, proportion explained = -1.2% [-2.9%, -0.1%], *p* = .016.

It might also be conjectured that the reason why the above findings do not support the theory is that List 5 interim test recall in the Restudy group was at floor (*M* = 2.65). To test this assumption, for each of the Test and Restudy group, we only retained the top quartile of participants. Importantly, with this restriction, recall performance for the Restudy group was at 7.64, which should significantly allay any concern about a floor effect. We then repeated all the previous mediation analyses, but none of them provided evidence supporting the reset-of-encoding theory.

Supplemental analyses to compare the contributions of release-from-PI and temporal processing strategy-change

To mitigate the data deletion issues, List 5 TCSs were estimated using the linear interpolation method for the 303 participants for whom these scores were not computable, and we then re-ran the mediation analysis with all 1,032 participants’ data included. Again, the results showed a significant indirect effect through release-from-PI (1.054 [0.773, 1.343]) and a significant indirect effect through temporal processing strategy-change (0.482 [0.348, 0.634]). In addition, the mediating effect of release-from-PI was stronger than that of temporal processing strategy-change, difference in indirect effect = 0.572 [0.264, 0.890].

When instead the analysis was restricted to the top quartile of participants, the results again showed a significant indirect effect through release-from-PI (2.422 [1.811, 3.101]) and a significant indirect effect through temporal processing strategy-change (0.305 [0.096, 0.560]). In addition, the mediating effect of release-from-PI was stronger than that of temporal processing strategy-change, difference in indirect effect = 2.117 [1.423, 2.865].

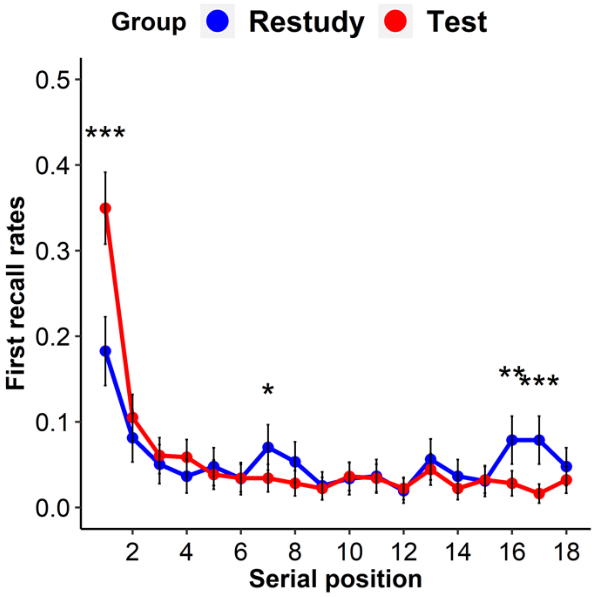

Figure A1: Proportions of participants who initiated List 5 recall at each serial position. Error bars represent 95% CI. ****p* < .001; ***p* < .01; **p* < .05.