Journal of Memory and Language 130 (2023) 104412

Contents lists available at ScienceDirect

Journal of Memory and Language

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

Evaluating the conceptual strategy change account of test-potentiated new learning in list recall

Shaun Boustani [a,*], Caleb Owens [b], Hilary J. Don [a], Chunliang Yang [c], David R. Shanks [a]

[a] University College London, United Kingdom
[b] University of Sydney, Australia
[c] Beijing Normal University, China

ARTICLE INFO

Keywords:
Test-potentiated new learning
Forward testing effect
Interpolated retrieval
Testing effect
Retrieval practice

ABSTRACT

Prior testing potentiates new learning, an effect known as test-potentiated new learning (TPNL). Research using lists of related words has established that testing, by free recall, also increases semantic clustering of later recall output. It has been suggested that this is evidence that testing induces a strategy change in encoding and retrieval towards greater conceptual organisation. The current research evaluated whether this conceptual strategy change explains TPNL in three experiments. We found a) that a retrieval task that did not increase semantic clustering (list discrimination) consistently produced TPNL, and b) that factors (word-relatedness and list structure) that influenced the amount of semantic clustering had no effect on the magnitude of TPNL. These results suggest that conceptual strategy change is neither necessary nor sufficient for TPNL and is more likely to be an effect of testing, rather than a cause of TPNL.

Introduction

Testing can enhance the learning of subsequently presented new information, an effect known as *test-potentiated new learning* (TPNL; Szpunar, McDermott, & Roediger, 2008). A recent meta-analysis has confirmed that it is a reliable and robust effect (Chan, Meissner, & Davis, 2018; but see Boustani & Shanks, 2022), with obvious relevance to classroom learning (Jing, Szpunar, & Schacter, 2016). However, the mechanisms underlying TPNL remain elusive.

In a classic demonstration of TPNL, Szpunar et al. (2008, Experiment 3) had participants study-five lists of interrelated words. After each of lists 1–4, participants in a retrieval group attempted to retrieve the words from the previous list, while those in a control group restudied the items on that list. After studying list 5, both groups were tested in a *criterial test*. The now well-replicated finding was that participants who had previously retrieved words recalled significantly more list 5 words than those who had not. Thus, testing can potentiate new learning. Several theories have been proposed to explain this benefit (for reviews, see Chan, Meissner, & Davis, 2018; Pastötter & Bäuml, 2014; Yang, Potts, & Shanks, 2018) although it is likely that TPNL is a multi-faceted effect with several underlying causes.

The conceptual organisation strategy change account

One contributing mechanism of TPNL that has received some support hypothesizes that testing may partially potentiate new learning by inducing a strategy change in encoding and retrieval, increasing the use of semantic categories. Chan, Manley, Davis, and Szpunar (2018) found that alongside potentiating the new learning of a word list, testing also increased the semantic clustering of recall output, relative to restudy. That is, participants who had been previously tested not only recalled more words from the final list than those in the restudy condition, they also tended to recall them in clusters organized according to their semantic categories. Chan et al. interpreted this as evidence that testing induced a change in encoding strategy whereby participants in the testing condition processed the newly learned materials according to how they were semantically related to other words within the list and to words from previous lists. This change allows for efficient encoding, as new materials are sorted into pre-existing categories, and supporting more efficient retrieval as the categories serve as effective retrieval cues during recall. However, whereas it is clear that testing results in a change to the clustering of recall output – indicative of a conceptual strategy change – the degree to which this is a causal mechanism of TPNL is not.

* Corresponding author.
E-mail address: s.boustani@ucl.ac.uk (S. Boustani).

https://doi.org/10.1016/j.jml.2023.104412
Received 6 March 2022; Received in revised form 18 January 2023; Accepted 31 January 2023
Available online 6 March 2023

There are two relevant details about Chan, Manley et al.'s (2018) study which should be noted. Firstly, they used lists of semantically interrelated words, and secondly, word order within each list was randomised. These factors are important because the use of interrelated word lists makes conceptual organisation more apparent and probably increases the efficacy of intra-list integration, and randomisation makes the associations between words less obvious. As such, increased semantic clustering following testing is indicative that testing highlights the semantic associations between words, which is argued to consequently potentiate new learning (Chan, Manley et al., 2018).

These factors define a conceptual strategy change which is 'intentional' or 'planned'. That is, the conceptual structure of the lists is created by the researcher and the associations that might be adopted by learners are prepared. Any strategy change is therefore limited to materials where the associations between items are evident, such as in interrelated word lists, the focus of this article. However, it is important to note that *ad hoc* strategy changes are also possible, that is, where the conceptual structure of lists is not prepared by the researcher, and instead learners form associations between words based on their idiosyncratic individual conceptualizations (Tulving, 1962). It has been argued that these *ad hoc* relations are more effortful to construct and that TPNL is likely to be weaker with unrelated lists:

"The effort required to produce these ad-hoc relations would likely be greater than that needed to process the pre-existing associations for semantically related words, so the TPNL effect should be smaller with unrelated word lists than moderately related word lists" (Chan, Manley et al., p.94).

In a recent study, Kliegl and Bäuml (2021) compared TPNL in related and unrelated word lists. They found that retrieval practice potentiated new learning in both cases, but a semantic generation task only potentiated new learning in unrelated lists. In addition, they found that TPNL for both list types was maintained at a list 4 study-retention interval of 1-min, and a list 4 lag of 1-min, but at a 25-min retention interval and 25-min lag was only maintained for related lists. These results suggest that TPNL is mainly driven by strategy change with categorized materials and by context change with unrelated materials.

To determine the causal influence of strategy change on TPNL, we aimed to investigate whether strategy change is necessary and/or sufficient for the effect to occur. One way to test whether conceptual strategy change is *necessary* is by using an episodic retrieval task which does not increase semantic clustering and observe whether, and to what extent, it potentiates new learning. In list discrimination tasks, participants are provided the words from a previous list and are asked to recall which list they are from. List discrimination has been used in testing effect research and has been demonstrated to produce less semantic clustering of recall output than restudy (Whiffen & Karpicke, 2017). However, whether list discrimination tasks influence clustering of newly learnt materials or potentiate new learning has not yet been assessed. A list discrimination task can be used to test the impact of a conceptual strategy change because it exclusively focuses participants on temporal cues and hence should not facilitate a strategy change towards conceptual organisation.

There is reason to suspect that list discrimination tasks will potentiate new learning. Research has demonstrated TPNL-like effects using other tasks, such as semantic generation (Divis & Benjamin, 2014) and the recall of autobiographical information (Pastötter, Bäuml, & Hanslmayr, 2008; Pastötter, Schicker, Niedernhuber, & Bäuml, 2011). Additionally, studies have demonstrated a benefit of context-reinstatement on learning within other paradigms, such as directed forgetting (Jonker, Seli, & MacLeod, 2013; Sahakyan & Kelley, 2002). These effects are thought to occur due to an internal context change, which could also explain how new learning is potentiated by testing (although there is debate regarding the role of context change when learning related materials; Kliegl & Bäuml, 2021). As such, determining whether list discrimination can produce a reliable and robust TPNL in word list learning would shed light on the potential role of strategy change following retrieval practice.

To test whether conceptual strategy change is a *sufficient* cause of TPNL, we manipulated conditions that should alter the magnitude of semantic clustering, in order to observe potential effects on the magnitude of TPNL. One prediction from Chan, Manley et al.'s (2018) theorizing is that TPNL should be larger when words within and between lists are interrelated, but smaller when they are not. Of course, it is important to note that much research has demonstrated robust and reliable TPNL effects in unrelated materials (Kliegl & Bäuml, 2021; Kliegl, Kriechbaum, & Bäuml, 2022; Wissman, Rawson, & Pyc, 2011), a finding which has been confirmed through meta-analysis (Chan, Meissner, & Davis, 2018). However, the magnitude of TPNL in related and unrelated word lists has not been directly compared. Of some relevance to this issue, Ahn and Chan (2022) directly compared TPNL in lists of words which either shared categorical relationships within and between lists or just within lists. In their Experiments 1 and 2, five lists were constructed from words taken from four taxonomic categories (fruits, animals, body parts, and sports). In the first four lists, each list was either composed of words taken from a single category (e.g., list 1 only contained fruits, list 2 animals, and so on), or words taken from each category (i.e., all lists contained words from all four categories), while the final list was always constructed from words taken from all categories. Ahn and Chan found that TPNL magnitude was approximately equivalent between the two groups. However, in their study, word lists in the unrelated condition still shared categorical relationships within lists. That is, it remains possible that a conceptual strategy change was still being used by participants but focused on the categorisation of words within a list. As such, it remains unclear whether the magnitude of TPNL will be equivalent when using materials that are unrelated both between and within lists. Given that semantic clustering is only possible when using related materials, finding that TPNL is equivalent would imply that conceptual strategy change is insufficient to cause TPNL.

A related prediction is that if differences in recall between restudy and retrieval are due to the semantic associations between words being made more obvious, then TPNL should be reduced when those associations are made inherently obvious to all participants. One way of making these associations salient is by presenting words in order according to the categories they belong to, rather than randomised order. As the category groupings are equally clear under such circumstances for participants in both the restudy and retrieval groups, there should be a reduction in the magnitude of TPNL. To that extent, the available research on the impact of list structure on new learning has been the subject of some debate. For example, in a cross-experiment comparison, Nunes and Weinstein (2012) found that structuring lists created from DRM words abolished TPNL, but in a replication with a much larger sample Ahn and Chan (2022) found that TPNL was present and unaffected by list categorisation, a finding they confirmed through a meta-analysis. However, in both cases, semantic clustering was not measured, and so the impact of structure on conceptual strategy usage is unknown.

The current study

The current study uses these assumptions to test the degree to which conceptual strategy change contributes to TPNL in word lists, and to provide a conceptual replication of the main findings of Chan, Manley et al. (2018).

Experiment 1 tests whether TPNL is moderated by the relatedness of words in a list, and explores whether list discrimination tasks potentiate new learning without increasing semantic clustering. It does so by comparing new learning in participants completing either free recall tests, list discrimination tests, or restudy, after studying word lists comprising highly-related words or words of mixed-relatedness. The word lists are either constructed from a small set of taxonomic categories, such that they are semantically associated (high-relatedness lists), or from a larger mixed set of categories with some words being

semantically associated within a list, and others not (mixed-relatedness lists). An important aspect of the mixed-relatedness word list is that it is constructed of three distinct types, a group of highly related words, a group of words with medium relatedness, and a group of unrelated words. As such, Experiment 1 also compares the magnitude of TPNL with words that are highly related, of medium relatedness, or unrelated within a list.

Experiment 2 manipulated the structure of the word lists. As discussed above, the strategy change account proposes that an important mechanism through which testing potentiates new learning relative to restudy is via making the conceptual organisation of lists more obvious. As presenting words according to their taxonomic categories should make the associations between them evident for those in both conditions, the difference in new learning following retrieval and restudy should be reduced. Analysing the semantic clustering of recall output, where there should also be no differences between conditions, will determine whether this is the case.

Experiment 3 serves as a pre-registered conceptual replication of Experiments 1 and 2 and, for reasons described below, incorporates methodological changes.

Experiment 1

Experiment 1 evaluates the potential role of a conceptual strategy change in three ways. The first is that it examines whether retrieval in the form of list discrimination produces TPNL without increasing semantic clustering, the second is that it compares TPNL magnitude between high-relatedness and mixed-relatedness word lists, and finally it examines whether TPNL magnitude is different for words of the mixed-relatedness list which are either highly related, of medium relatedness, or unrelated within the list. This experiment also serves as a conceptual replication of the 1-minute retention interval conditions from Chan, Manley et al. (2018), with a different set of materials and a different participant sample.

In Experiment 1, participants studied three lists of words, each followed by a distractor task, and then an interim task (see Fig. 1 for a schematic diagram of the design). Word lists were composed of either highly related words (words which were related to others both within and between lists; the high-relatedness word lists) or a combination of highly related words, words of medium relatedness, and unrelated words (the mixed-relatedness word lists). The word lists for all experiments, including the mixed-relatedness word lists, are shown in Tables S1, S7, and S8 in the Supplementary Information. The first two interim tasks were dependent on which condition the participant was allocated to: either free recall of words from the previous list, a list discrimination task in which participants make a source judgment about which list a word was previously presented in, or restudy of the previous list. The final list was classified as new learning and the final interim task was free recall (the criterial test).

The strategy change account makes clear predictions about the semantic clustering of recall output and TPNL magnitude. If conceptual strategy change plays a major role in TPNL, then testing should result in greater semantic clustering of output and TPNL should be observed for high-relatedness word lists. However, as semantic associations should be less likely when using mixed-relatedness word lists, semantic clustering of output should be low and TPNL magnitude substantially reduced when using such lists.

Likewise, if list discrimination results in weaker semantic clustering of output, as would be predicted based on the findings of Whiffen and Karpicke (2017), then list discrimination – despite being a retrieval task – should potentiate new learning less effectively than free recall tests for both high-relatedness and mixed-relatedness word lists.

The mixed-relatedness word list also allows for within-group analysis, useful for further testing the role of conceptual strategy change in TPNL. The account predicts the largest TPNL effect for highly-related words, a smaller TPNL effect for medium-relatedness words, and the smallest effect for unrelated words. The strategy change account predicts that recall of highly related words on the critical final list (List 3) should be strongly facilitated by previous testing (compared to restudy), because testing will encourage encoding of words according to their semantic categories. This is not possible for unrelated words, and so recall should be substantially reduced. Mixed-relatedness words should fall in between these extremes.

In this and the subsequent experiments we report how we determined our sample size, all data exclusions (if any), all manipulations, and all measures.

The data that support the findings of this and the subsequent studies are available from the Open Science Framework (OSF): https://osf.io/a25m9. All analyses were conducted in SPSS, with Bayes factors calculated using JASP.

Method

Participants

In total, 317 first year psychology students (*M* age = 19.91, *SD* = 4.51; 108 male, 207 female, 2 undeclared) participated in the study as part of an introductory tutorial. Classes comprised approximately 15–22 students. Two hundred and thirty-four indicated that English was their first language. In the conditions using mixed-relatedness word lists, 67 completed list discrimination, 76 completed restudy, and 74 completed free recall. In the conditions using high-relatedness word lists, 32 completed list discrimination, 34 completed restudy, and 34 completed free recall. The greater allocation of participants to the mixed-relatedness conditions increased the power of the within-subjects analyses described below. Although sample size was determined by class

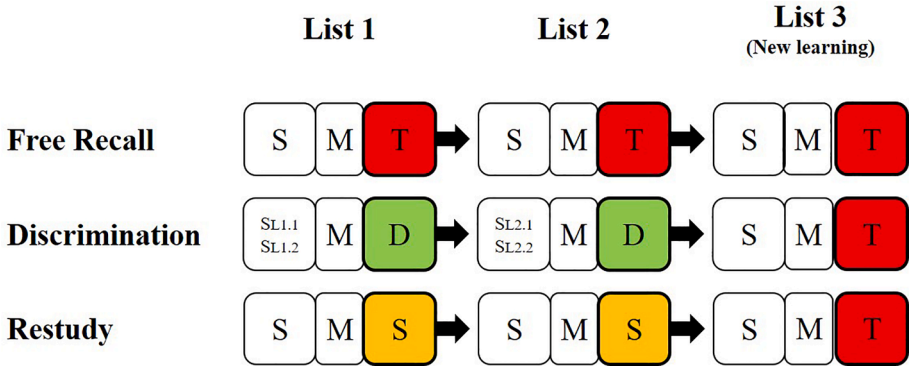

Fig. 1. Schematic diagram of procedure for Experiment 1. Participants first studied two lists of words and then completed different interim tasks. Study of list 3 was always followed by free recall, which served as the criterial test for new learning. S refers to study, M to the distractor task (mental arithmetic), and T (Recall Test) and D (Discrimination) refer to the corresponding interim tasks. Study in list discrimination was split into two lists to allow for source discrimination.

size, we conducted a post-hoc sensitivity analysis using G*Power 3.1 (Faul, Erdfelder, Buchner, & Lang, 2009) which indicated that our total sample of 317 was sufficient to detect a small main effect ($f = 0.18$) and interaction with $\alpha = .05$ and power $= .80$ in a 3 × 2 factorial analysis of variance (ANOVA).

Materials

For the high-relatedness word lists, 36 words were selected from the Van Overschelde, Rawson, and Dunlosky (2004) category norms, from which three interrelated lists of 12 words were constructed. The nine highest-frequency exemplars were selected from each of four taxonomic categories (animals, fruits, musical instruments, weather). These were then allocated to three lists, so that each list included three exemplars from each of the four categories. Thus, each word had 2 taxonomically-related associates in the same list and 3 associates in each of the other lists. Word order within a list was randomised, and list order was counterbalanced between-subjects using a Latin square (i.e., four different orders where each list occurred once at each sequence position).

For the mixed-relatedness word lists, another set of 36 words was selected from the Van Overschelde et al. (2004) norms. The twelve highly related words were the 6 highest-frequency words from two taxonomic categories. The twelve medium relatedness words were the three lowest-frequency words from four categories. The 12 unrelated words were the highest-frequency words from 12 unrelated categories. The words used to create this list are shown in Table S1 (Supplementary Information). Each highly-related word had taxonomically-related associates both within and across lists; each medium-related word had taxonomically-related associates across but not within lists; and each unrelated word was taxonomically unique. It is important to note that, in order to meet the design constraints, different words were used for the related and unrelated conditions. To assess possible differences between word lists, we compared the lexical characteristics of the related and unrelated word lists on a range of factors known to be important to recall (Lau, Goh, & Yap, 2018: specifically, word length, word frequency, concreteness, age of acquisition, and semantic and emotional factors). Means and significance tests are reported in Table S2 within the Supplementary Information, but in summary we found no significant differences.

Design

The experiment utilised a 3 (Interim task: Free Recall, List Discrimination, Restudy) × 2 (Relatedness: High-Relatedness Lists, Mixed-Relatedness Lists) between-subjects design.

Procedure

For all experiments, ethical approval was provided by the University of Sydney ethics committee (HREC). Prior to beginning the experiment, consent was obtained and participants answered a demographic questionnaire. Each participant completed the study on an individual computer in a classroom. At the beginning of the experiment participants were informed that they would study several word lists and that after each list they would complete a task. Following instructions, participants attempted a trial of their interim task and then began the experiment.

Participants in every condition studied a list of words, completed a distractor task, and then an interim task. For Lists 1 and 2 participants completed a restudy, list discrimination, or free recall task. After studying List 3 all participants completed free recall.

During study, the list number (i.e., List 1, List 2, List 3) was presented in the middle of the screen for 2 s followed by a fixation cross for 2 s. After this, words appeared on the screen one at a time for 3 s followed by a 500 ms interstimulus interval, progressing automatically. The subsequent distractor task was 40 s of simple arithmetic problems.

Restudy was a repeat of study. During free recall, participants were instructed to retrieve all words from the previous list. Words were typed and remained on screen. The test lasted 60 s before automatically progressing. For list discrimination, the study list was split in half with the heading 'List X.1' at the beginning, and then after half of the words were presented the heading changed to 'List X.2' (e.g., List 1.1 and List 1.2). This was done so list discrimination was possible from List 1. The list discrimination task involved displaying a study word on the screen for 3 s during which participants were asked to indicate whether that word was from List X.1 or List X.2 by clicking on labelled buttons under the word. The words remained on screen for 3 s regardless of when participants made their responses and the computer program automatically advanced to the next word even if a response had not been made. There was a 500 ms interstimulus interval between items.

Following the criterial recall test on List 3, participants completed a 60 s distractor task and a cumulative test which required free recall of all lists for 120 s. The results of this test, and the equivalent cumulative tests in subsequent experiments, are reported in the Supplementary Information (Table S6).

Semantic clustering of output was measured using the same formula as Chan, Manley et al. (2018), the adjusted-ratio-of-clustering method (ARC: Roenker, Thompson, & Brown, 1971). Scores can range from negative values to 1, with higher ARC scores indicating greater semantic clustering, and 0 indicating chance level clustering (negative values indicate that participants are recalling words in a manner actively opposing conceptual clustering).

Results

For non-significant findings, we report Bayes factors favouring the null (BF_{01}). For the reader's benefit, we include a summary of the main qualitative results for this and the subsequent experiments in Table 1.

Recall

A 3 × 2 ANOVA was conducted analysing the impact of interim task (free recall, list discrimination, restudy) and content relatedness (high vs mixed) on List 3 recall. A significant main effect of interim task was

Table 1
Summary of major findings from Experiments 1–3.

Measure	FR vs RS	FR vs LD	LD vs RS	High vs Mixed	Struc vs Rand
Total Recall					
Experiment 1	✔	✔	✔	✔	
Experiment 2	✔				
Experiment 3	✔	✔	✔		✔
Intrusions					
Experiment 1	✔	✔	✔	✔	
Experiment 2	✔				
Experiment 3	✔	✔	✔		✔
Semantic clustering					
Experiment 1	✔	×	×	✔	
Experiment 2	✔				
Experiment 3	✔	✔	×		✔

Note. FR = Free Recall, RS = Restudy, LD = List Discrimination, High = Highly related lists, Mixed = Mixed relatedness lists, Struc = Structured lists, Rand = Randomised lists. Ticks denote significant differences ($p < .05$), crosses denote non-significant differences.

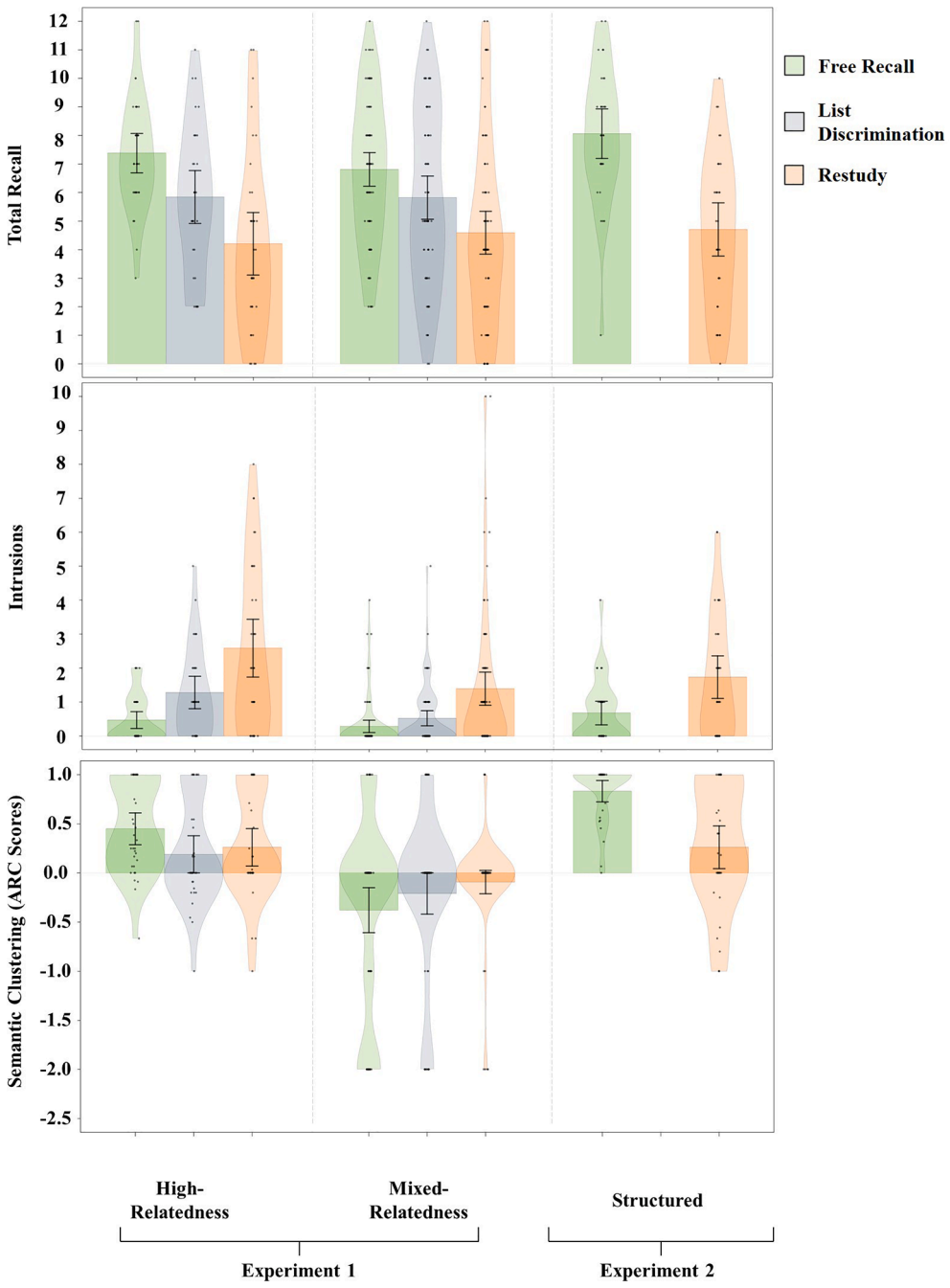

Fig. 2. Impact of material type and interim task in Experiments 1 and 2. High-relatedness and mixed-relatedness materials were used in Experiment 1 and refer to the relationship between words. Structured materials, in which words were presented in their taxonomic categories, were used in Experiment 2. The horizontal line within each distribution represents the group mean and error bars represent 95% CI. Upper panel: Recall of List 3 words. Middle panel: Prior list intrusions in List 3 test. Lower panel: Semantic clustering of List 3 recall, measured by ARC scores.

found, $F(2, 311) = 20.60$, $p < .001$, $\eta_p^2 = .117$. However, neither a main effect of relatedness, $F(1, 311) = 0.04$, $p = .842$, $BF_{01} = 10.20$, nor an interaction, $F(2, 311) = 0.652$, $p = .521$, $BF_{01} = 19.04$, was observed. See the upper panel of Fig. 2 for a violin plot of the data.[1]

Three planned contrasts were performed to compare recall in the free recall, list discrimination, and restudy groups with Holm-Bonferroni adjustments to control for multiple comparisons (described p-values have been adjusted). Those contrasts revealed that the free recall group ($M = 6.99$, $SD = .165$) achieved significantly greater recall than both the list discrimination ($M = 5.83$, $SD = 2.93$), $t(205) = 3.14$, $p = .004$, $d = 0.44$, and restudy ($M = 4.47$, $SD = 3.23$), $t(216) = 6.54$, $p < .001$, $d = 0.89$, groups, and the list discrimination group also achieved significantly greater recall than the restudy group, $t(207) = 3.17$, $p = .002$, $d = 0.44$. These results indicate that both free recall and list discrimination produce TPNL, but that the effect is larger in free recall.

Intrusions

A 3 (interim task) × 2 (relatedness) ANOVA was also conducted on prior list intrusions. A main effect of interim task was found, $F(2, 311) = 27.80$, $p < .001$, $\eta_p^2 = .152$. High-relatedness materials ($M = 1.45$, $SD = 1.87$) produced more prior list intrusions than mixed-relatedness materials ($M = .75$, $SD = 1.51$), $F(1, 311) = 15.24$, $p < .001$, $\eta_p^2 = .047$. The interaction was not significant, $F(2, 311) = 2.61$, $p = .075$, $BF_{01} = 0.51$. See Fig. 2 (middle panel) for a violin plot of the data.

Planned contrasts revealed that participants in the free recall group

[1] For completeness, we report total recall, semantic clustering and intrusions during the free recall interim tasks in Experiments 1 and 2 (Table S4) and Experiment 3 (Table S5) in the Supplementary Information.

(*M* =.34, *SD* =.76) recalled significantly fewer intrusions from prior lists than those in either the list discrimination (*M* =.77, *SD* = 1.11), *t*(205) = 3.23, *p* =.001, *d* = 0.45, or restudy (*M* = 1.76, *SD* = 2.29), *t*(216) = 6.12, *p* =.0001, *d* = 0.83, groups, and that participants in the list discrimination group also recalled significantly fewer intrusions than those in the restudy group, *t*(207) = 3.93, *p* =.0003, *d* = 0.54.[2]

Semantic clustering

A final 3 (interim task) × 2 (relatedness) ANOVA was conducted on semantic clustering of recall output (ARC scores). No main effect of interim task was found, *F*(2, 311) = 0.38, *p* =.686, BF_{01} = 8.98. However, high-relatedness materials (*M* =.30, *SD* =.52) resulted in greater semantic clustering than mixed-relatedness materials (*M* = -.23, *SD* =.82), *F*(1, 311) = 35.47, *p* <.001, η_p^2 =.102. The interaction was not statistically significant, *F*(2, 311) = 2.97, *p* =.053, BF_{01} = 3.40. See Fig. 2 (lower panel) for a violin plot of the data.

Because very little clustering was possible in the mixed-relatedness conditions, we conducted restricted simple effects comparisons analysing ARC scores for the high-relatedness lists only, with Holm-Bonferroni adjustments. This analysis revealed that recall output clustering in the free recall group (*M* =.45, *SD* =.46) tended to be larger than in the list discrimination group (*M* =.19, *SD* =.52), *t*(64) = 2.14, *p* =.11, *d* = 0.53, B_{01} = 0.58, and in the restudy group (*M* =.26, *SD* =.55), *t*(66) = 1.53, *p* =.26, B_{01} = 1.48. The difference in semantic clustering between the list discrimination and restudy groups was negligible, *t*(64) = 0.55, *p* =.586, BF_{01} = 3.49.

These results qualitatively replicate the major finding from Chan, Manley et al. (2018), indicating that semantic clustering of recall output is numerically greater following free recall than restudy, although the effect was not statistically significant in our sample.

Within-subjects analysis

The conceptual strategy change account predicts the largest TPNL effect for highly-related words, a smaller TPNL effect for medium-relatedness words, and the smallest effect for unrelated words. We therefore compared final list recall in the free recall and restudy groups between-subjects, and for the three relatedness types within-subjects. We dropped list discrimination from this comparison as it was not relevant to the current issue and to restrict the number of comparisons. Means for list discrimination by relatedness can be found in Table S3 (Supplementary Information) and are not further discussed.

Recall. A two-way mixed ANOVA revealed a main effect of interim task, with the free recall group (*M* = 2.27, *SD* = 1.14) recalling significantly more List 3 words than the restudy group (*M* = 1.24, *SD* = 1.31), *F*(1, 148) = 21.38, *p* <.001, η_p^2 =.126. A main effect of relatedness was also found, *F*(2, 296) = 7.79, *p* =.001, η_p^2 =.05. The interaction was not significant, *F*(2, 296) = 2.46, *p* =.087, BF_{01} = 0.65. See Fig. 3.

Post hoc analyses compared recall for different levels of content relatedness, with Holm-Bonferroni adjustments. Recall of highly-related words (*M* = 2.11, *SD* = 1.25) was greater than for medium-related words (*M* = 1.70, *SD* = 1.27), *t*(149) = 3.94, *p* <.001, *d* = 0.32, and marginally better than for unrelated words (*M* = 1.88, *SD* = 1.32), *t*(149) = 2.19, *p* =.059, BF_{01} = 1.28, *d* = 0.18. Recall of medium-relatedness words was marginally lower than that of unrelated words, *t*(149) = 1.75, *p* =.081, BF_{01} = 2.87, *d* = 0.14.

Intrusions. A similar analysis showed a main effect of interim task, with the free recall group (*M* = 0.11, *SD* =.38) experiencing substantially fewer intrusions than the restudy group (*M* =.68, *SD* =.85), *F*(1, 148) = 15.86, *p* <.001, η_p^2 =.097. A main effect of relatedness was also found, *F*(2, 296) = 7.67, *p* < .001, η_p^2 =.049. The interaction was not significant, *F*(2, 296) = 1.05, *p* =.351, BF_{01} = 2.27. See Fig. 3.

Post hoc analyses with Holm-Bonferroni adjustments revealed more intrusions in the recall of highly-related words (*M* =.41, *SD* =.82) compared to both medium-related (*M* =.24, *SD* =.65), *t*(149) = 3.19, *p* =.003, *d* = 0.26, and unrelated words (*M* =.22, *SD* =.55), *t*(149) = 3.56, *p* =.001, *d* = 0.29. The difference in prior list intrusions between the medium-related and unrelated subsets was not significant, *t*(149) =.37, *p* =.713, BF_{01} = 10.28. As in the between-subjects comparison described above, these analyses indicate that highly-related materials are associated with increased prior list intrusions.

Discussion

The purpose of Experiment 1 was to test the conceptual strategy change account of TPNL in word lists by assessing whether the magnitude of TPNL is modulated by word list relatedness and whether a list discrimination task, which we hypothesized would not result in greater semantic clustering of recall output than restudy, would produce a smaller TPNL effect.

We reproduced the major findings of Chan, Manley et al. (2018). TPNL was produced when using a free recall task (with a large effect size of almost 0.90), and the pattern of results demonstrated that free recall may potentiate semantic clustering of recall output, relative to restudy.[3] However, we did not find good evidence for a causal role of conceptual strategy change in TPNL.

We observed TPNL with two retrieval tasks, a free recall task and a list discrimination task that does not encourage conceptual strategy change. List discrimination did result in weaker TPNL than free recall, and this difference was proportionate to the difference in semantic clustering. This result may suggest that the difference in recall is due to a difference in semantic clustering. However, list discrimination improved recall relative to restudy without any commensurate increase in semantic clustering, thus providing clear evidence that conceptual strategy change is not necessary for TPNL to occur. In addition, the strategy change account predicted that TPNL would be reduced when using mixed-relatedness word lists and for unrelated words within those lists. We instead found that, despite reliably less semantic clustering observed when using mixed-relatedness word lists, TPNL was produced when using both high-relatedness and mixed-relatedness word lists, and to a similar magnitude. Additionally, TPNL was not substantially different for words that were highly-related, of medium relatedness, or unrelated, within a list. These results confirm the trends discussed above which suggested that TPNL was large and reliable when using unrelated materials. It also suggests that increasing semantic clustering (by way of using highly-related word lists), is not sufficient to increase TPNL. In sum, these results are important because they suggest that although semantic clustering was increased by free recall, this is unlikely to be the causal mechanism behind the observed TPNL. This is further examined in Experiment 2.

Experiment 2

Experiment 2 assessed the conceptual change account by using the same high-relatedness materials as in Experiment 1, but presenting the words according to their taxonomic categories, rather than in random order. The strategy change account hypothesizes that testing makes the conceptual connections between words more obvious than restudy. One

[2] A mediation analysis is presented in the Supplementary Information which explores the possible role of conceptual strategy change in accounting for the difference in prior list intrusions between groups studying high-relatedness and mixed-relatedness word lists.

[3] Although this was numerically true, it is important to note that for the direct replication of Chan, Manley et al.'s (2018) 1-minute retention interval conditions (that is, the high-relatedness groups comparing free recall and restudy), the difference in semantic clustering did not achieve statistical significance. To foreshadow, the equivalent contrast in Experiment 3 is statistically significant.

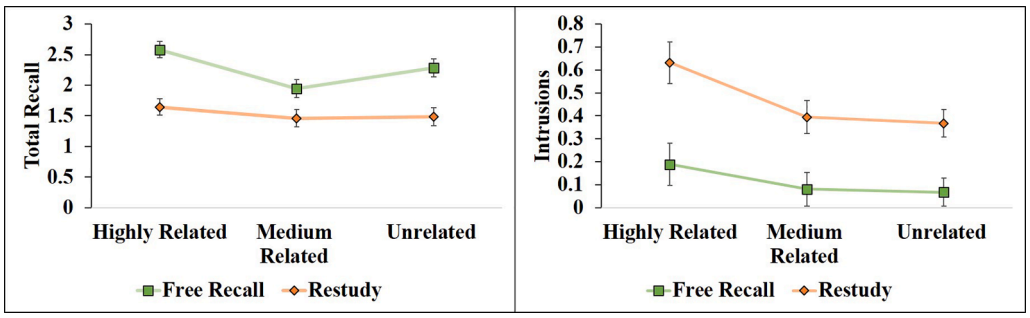

Fig. 3. Left panel. Mean recall of List 3 words for participants presented with mixed-relatedness lists as a function of within-list material relatedness and interim task. **Right panel.** Mean prior list intrusions as a function of material relatedness and interim task. Markers represent means of each condition. Error bars represent SE.

way to test this assumption is to present the word list organized by taxonomic category, so as to render the conceptual associations evident for all groups. If TPNL is reduced by presenting the words according to their categories, then the strategy change account is supported.

Method

Participants

In total, 65 first year psychology students (M age = 20.14, SD = 4.57; 25 male, 40 female) participated in the study as part of an introductory psychology tutorial, and data collection occurred simultaneously with Experiment 1. Classes comprised approximately 15–22 students. Forty-eight indicated that English was their first language. Thirty-one participants completed free recall and 34 restudy. Sample size was determined by class size, but a power analysis using G*Power 3.1 (Faul et al., 2009) indicated that a minimum total sample size of 42 was adequate to detect a moderate-large main effect (d = 0.70) with α =.05, and power =.80 in an independent samples t-test. The effect size was again based on the results of Chan, Manley et al. (2018).

Materials

Materials were the same 36 words from the high-relatedness lists used in Experiment 1.

Procedure

The procedure of Experiment 2 was identical to Experiment 1 except that rather than word order being randomised, words were presented in order according to their taxonomic categories. That is, in each list, successive groups of 3 words all came from the same category, and the order was fixed and the same for all participants. No list discrimination task was included.

Results

Recall

An independent samples t-test revealed that participants in the free recall group (M = 8.06, SD = 2.37) recalled significantly more words than those in the restudy group (M = 4.71, SD = 2.67), $t(63)$ = 5.35, p <.001, d = 1.33 (see Fig. 2, upper panel, for a violin plot of the results).

Intrusions

An independent samples t-test revealed that the free recall group (M =.68, SD =.94) generated significantly fewer prior-list intrusions than the restudy group (M = 1.74, SD = 1.80), $t(63)$ = 2.93, p =.005, d = 0.73. See Fig. 2 (middle panel) for a violin plot of the results.

Semantic clustering

An independent samples t-test revealed that the semantic clustering of recall output in the free recall group (M =.83, SD =.30) was significantly greater than in the restudy group (M =.26, SD =.62), $t(63)$ = 4.65, p <.001, d = 1.16. This is surprising and goes against the hypothesis that structured lists would eliminate the difference in clustering. See Fig. 2, lower panel, for a violin plot of the data.

Discussion

The purpose of Experiment 2 was to assess whether TPNL would be reduced by presenting words structured according to their taxonomic categories, which we predicted would reduce the difference in semantic clustering between free recall and restudy groups, as it would make conceptual ties between words more obvious for participants (Chan, Manley et al., 2018). However, free recall produced a robust and reliable TPNL effect in structured lists, comparable in magnitude to that observed when using randomised word lists in Experiment 1.

Structuring lists did not have the effect on ARC scores that we hypothesized. Rather than reducing the difference in clustering between free recall and restudy, we instead observed a greater difference between free recall and restudy.[4] This is a novel result which has not been previously observed. Comparing the results of Experiments 1 and 2, retrieval practice appears to increase semantic clustering even when the semantic categories should be obvious to all participants. However most notably, increasing semantic clustering does not appear to translate into an improvement in recall. As an exploratory analysis, we compared ARC scores and recall in the free recall condition between the highly related lists across experiments (see Fig. 2). This showed that structured lists (M =.83, SD =.30) significantly increased semantic clustering compared to random lists (M =.45, SD =.46), $t(63)$ = 3.94, p <.001, d = 0.98, but did not significantly increase recall in the criterial tests, ($M_{structured}$ = 7.38, $SD_{structured}$ = 1.99 vs M_{random} = 8.07, SD_{random} = 2.37), $t(63)$ = 1.26, p =.21, BF_{01} = 2.01.

Given that this is a cross-experiment analysis, and that this difference was not anticipated, it would be premature to draw any strong conclusions about the role of strategy change. However, this finding, in addition to the results from Experiment 1, could imply that greater adoption of a conceptual strategy is insufficient to facilitate new learning to a greater degree. As such, we conducted a replication in Experiment 3.

Experiment 3

To recap, the results thus far are challenging for the conceptual strategy account as a major cause of TPNL. Experiment 1 established

[4] A meta-analysis examining the differences in effect size for all measures is included in the Supplementary Information.

that, although we observed a difference in the magnitude of TPNL between free recall and list discrimination, list discrimination tasks produced TPNL without increasing semantic clustering. Moreover, TPNL magnitude was largely unaffected by the degree of relatedness of words within and between lists. Experiment 2 established that presenting words according to their taxonomic categories did not eliminate the semantic clustering advantage of free recall compared to restudy, nor did it reduce TPNL. Furthermore, a comparison of results from Experiments 1 and 2 indicated a potential dissociation between semantic clustering and recall. Despite structured word lists increasing semantic clustering in the free recall group, this was not accompanied by an equivalent increase in criterial test recall. However, this dissociation relies on a cross-experiment comparison, and therefore should be confirmed within a single experiment.

In addition, we made changes to the procedure to ensure sufficient sensitivity to observe differences in the adoption of a changed conceptual strategy. As noted, although the difference in semantic clustering between the free recall and restudy groups was qualitatively identical to that found by Chan, Manley et al. (2018), the contrasts failed to reach significance. In Experiments 1 and 2, our use of 3 lists of 12 words meant that participants learned fewer words, made fewer total retrieval attempts, and took fewer tests than those in Chan, Manley et al.'s (2018) experiments (who learned 4 lists of 15 words). With more and longer lists there is greater potential for participants to learn the taxonomic structure of the word lists and more opportunity to induce and apply conceptual strategies, enhancing learning and recall. This could lead to considerably increased semantically clustered recall output by the final list, relative to that observed in the previous experiments. If the lists used in Experiments 1 and 2 were too short to allow for maximal semantic clustering to be observed, then the impact of a conceptual strategy change could have been masked.

Experiment 3 was pre-registered (https://osf.io/zx93q) and served as a conceptual replication and extension of Experiments 1 and 2. Here we used a between-subjects design manipulating both interim tasks (free recall, list discrimination, and restudy) and word order (randomised and structured), using high-relatedness materials.

Method

Participants

A power analysis using G*Power 3.1 (Faul et al., 2009) indicated that a minimum total sample size of 301 was needed to detect a small-medium main effect ($f = 0.18$) and interaction with $\alpha = .05$ and power = .80 in a 3 × 2 factorial analysis of variance (ANOVA). This effect size was chosen as a mediation analysis[5] based on Experiments 1 and 2 indicated that the role of conceptual strategy change in TPNL is small in magnitude at most. In total, 312 undergraduate psychology students from the University of Sydney (M age = 20.26, SD = 2.70; 86 male, 222 female, 4 undeclared) participated in return for course credit. One hundred and seventy-one indicated that English was their first language. In the randomised word list conditions, 48 completed list discrimination, 53 completed restudy, and 53 completed free recall. In the structured word list conditions, 53 completed list discrimination, 53 completed restudy, and 52 completed free recall.

Materials

To more closely follow the method used by Chan, Manley et al. (2018), we created four new lists of 16 words.[6] Each list contained four exemplars from four categories, taken from Van Overschelde et al. (2004). The four categories were precious stones, relatives, reading materials, and animals. The average taxonomic frequencies were significantly different between the five categories (M_{stones} = .25, SD = .31, $M_{relatives}$ = .36, SD = .27, $M_{reading}$ = .27, SD = .27, $M_{animals}$ = .55, SD = .35, $F(3, 64) = 3.25$, p = .028). However, they did not differ across the four lists (range = .31–.42), $F(3, 64) = 0.38$, p = .77, $BF_{01} = 8.09$.

Experiment 1 only split the lists in half in the list discrimination groups, and not in the other groups. To avoid this condition-specific feature, all groups were presented a study list which was split in half with the heading 'List X.1' at the beginning, and then, after half of the words were presented, the heading 'List X.2' (e.g., List 1.1 and List 1.2). For the randomised groups, word order within the whole list was randomised. For the structured groups, words were presented in their taxonomic categories, with category order counterbalanced using a Latin square and words within categories randomised. In both groups, list order was counterbalanced using a Latin square.

Design

The experiment employed a 3 (Interim task: Free Recall, List Discrimination, Restudy) × 2 (Word order: Randomised, structured) between-subjects design.

Procedure

The general procedure was similar to Experiments 1 and 2, with the exception that the study was run online. Half the participants were presented words in each list in order of taxonomic category, and half were presented words in a random order.

During study, the list number (i.e., List 1.1, List 2.1, List 3.1, List 4.1) was presented in the middle of the screen for 2 s followed by a fixation cross for 1 s. After this, half the words appeared on the screen one at a time for 4 s followed by a 500 ms interstimulus interval, progressing automatically. The second list number (i.e., List 1.2, List 2.2, List 3.2, List 4.2) was presented followed by the other half of the words using the same procedure. The subsequent distractor task was 60 s of simple arithmetic.

Other than the described differences, procedures for the interim tasks were as in Experiments 1 and 2.

Following the final interim task, participants completed a 60 s distractor task and a cumulative test which required free recall of all lists for 120 s. The results of the cumulative test are reported in the Supplementary Information (Table S6).

Results

The experiment proceeded according to the pre-registration.

Recall

Fig. 4 (upper panel) shows a violin plot of the data. A 3 (interim task: recall, list discrimination, restudy) × 2 (word order: randomised vs structured) ANOVA was conducted on List 4 recall. As anticipated, a significant main effect of interim task was found, $F(2, 306) = 34.56$, p < .001, η_p^2 = .184. Neither a main effect of word order, $F(1, 306) = 0.51$, p = .477, $BF_{01} = 9.15$, nor an interaction, $F(2, 306) = 0.03$, p = .968, $BF_{01} = 26.21$, was observed.

Three planned contrasts with Holm-Bonferroni adjustments revealed that the free recall group ($M = 7.82$, $SD = 2.93$) achieved significantly greater recall than both the list discrimination ($M = 5.06$, $SD = 3.67$), t

[5] The results of this mediation analysis and a figure (Figure S2) demonstrating the model can be found in the Supplementary Information. We later present an updated model with data from all three experiments.

[6] While Chan et al. (2018) used 15 words per list, the inclusion of the List Discrimination group required an even number to allow for even splitting.

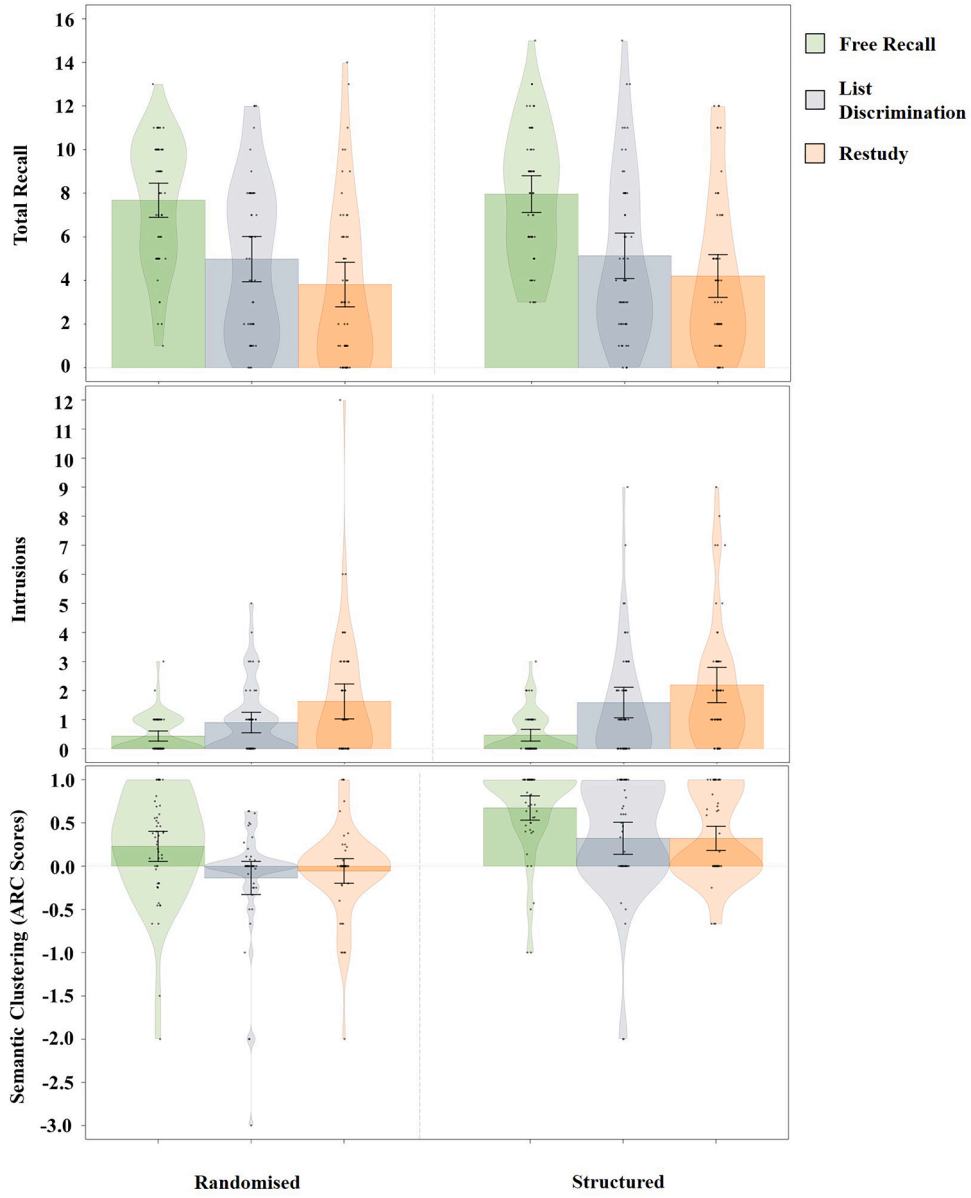

Fig. 4. Impact of material type and interim task in Experiment 3. The horizontal line within each distribution represents the group mean and error bars represent 95% CI. Upper panel: Mean recall of List 4 words. Middle panel: Mean prior list intrusions. Lower panel: Mean semantic clustering, measured by ARC scores.

(204) = 5.98, p <.001, d = 0.83, and restudy (M = 4.01, SD = 3.62), t (209) = 8.39, p <.001, d = 1.16, groups, and the list discrimination group also achieved significantly greater recall than the restudy group, t (205) = 2.07, p =.040, d = 0.29.

These results replicate the findings of Experiment 1 and confirm that both free recall and list discrimination produce TPNL, but that the effect is larger in free recall. They also replicate the earlier cross-experiment finding (see top panel of Fig. 2 and the meta-analysis reported in the Supplementary Information) that TPNL is approximately equivalent in randomised and structured lists.

Intrusions

Fig. 4 (middle panel) shows a violin plot of the intrusion data. A 3 (interim task) × 2 (word order) ANOVA found a main effect of interim task, F(2, 306) = 21.39, p <.001, η_p^2 =.123. Unexpectedly, a main effect of word order was also found with structured presentation (M = 1.42, SD = 1.87) producing more prior list intrusions than randomised presentation (M = 0.99, SD = 1.56), F(1, 306) = 5.42, p =.021, η_p^2 =.017. However, as predicted, the interaction was not significant, F(5, 306) = 1.22, p =.296, BF_{01} = 2.32.

Planned contrasts showed that participants in the free recall group (M =.45, SD =.68) generated significantly fewer intrusion responses than those in either the list discrimination (M = 1.26, SD = 1.64), t(204) = 4.66, p <.001, d = 0.65, or restudy (M = 1.91, SD = 2.20), t(209) = 6.50, p <.001, d = 0.89, groups, and that participants in the list discrimination group also generated significantly fewer intrusions than those in the restudy group, t(205) = 2.40, p =.017, d = 0.33.

Semantic clustering

See Fig. 4 (lower panel) for a violin plot of the data. A final 3 (interim task) × 2 (word order) ANOVA was conducted on semantic clustering of recall output (ARC scores). As anticipated, a main effect of interim task was found, F(2, 306) = 11.74, p <.001, η_p^2 =.071. As is very clear from the data in Fig. 4, a main effect of word order was also found with structured lists (M =.44, SD =.59) resulting in greater semantic clustering than randomised lists (M =.02, SD =.62), F(1, 306) = 41.46, p

<.001, η_p^2 =.119. However, the interaction was not significant, $F(5, 306) = 0.14$, p =.867, $BF_{01} = 3.47$.

Planned contrasts revealed that the semantic clustering of recall output in the free recall group (M =.45, SD =.61) was significantly greater than in the list discrimination group (M =.10, SD =.70), $t(204) = 3.77$, p <.001, $d = 0.526$, and greater than in the restudy group (M =.14, SD =.54), $t(209) = 3.98$, p <.001, $d = 0.55$. Importantly, clustering in the list discrimination group was not significantly different from that in the restudy group, $t(205) = 0.33$, p =.740, $BF_{01} = 6.11$.

The finding that semantic clustering of recall output is greater in the free recall group than in the restudy group confirms the trend observed in Experiment 1. In contrast, there was no evidence that semantic clustering was different between the list discrimination and restudy groups. This, in conjunction with the similar results from Experiment 1, indicates that list discrimination – despite inducing TPNL – does not foster a conceptual strategy change.

Discussion

Experiment 3 replicated the major results of the previous experiments and of Chan, Manley et al. (2018). Free recall resulted in substantially greater final list recall, greater semantic clustering of recall output and fewer prior list intrusions than either restudy or list discrimination. However, Experiment 3 also confirmed two complementary results that challenge the conceptual strategy change account. First, list discrimination tests can produce TPNL without increasing semantic clustering (as also found in Experiment 1). Secondly, increasing semantic clustering by presenting words according to their taxonomic categories did not improve recall in the final test (as found in the earlier comparison across Experiments 1 and 2).

Experiment 3 also explored the potential effect of word order on semantic clustering. The results confirmed both that the difference in semantic clustering between groups persisted when using structured word orders, and that the magnitude of this difference was approximately equal for randomised and structured lists, with a Bayesian analysis moderately supporting the null hypothesis. The corresponding implications for the conceptual strategy change account are considered in the General Discussion.

As in Experiment 2, Experiment 3 found that the difference in semantic clustering between the free recall and restudy groups remained significant when using structured list orders. Thus, the notion that presenting words according to their taxonomic categories would make conceptual associations equally evident for all conditions was not supported.[7] Instead, Experiment 3 demonstrated that structured word lists increase the use of conceptual strategies as evident in semantic clustering for all conditions, with a large effect size (η_p^2 =.118). Despite this, and contrary to the strategy change account, TPNL itself was equivalent in structured and randomised lists (effect size η_p^2 <.001). As in the comparison between Experiments 1 and 2, comparing ARC scores and recall in the free recall condition showed that learning structured lists (M =.67, SD =.50) significantly increased semantic clustering compared to random lists (M =.23, SD =.63), $t(103) = 3.99$, p <.001, $d = 0.78$, but did not significantly increase recall in the criterial tests, ($M_{structured}$ = 7.96, SD = 3.03 vs M_{random} = 7.68, SD = 2.85), $t(103) = 0.49$, p =.62, BF_{01} = 4.75 (see Fig. 4). This confirms the dissociation between semantic clustering and free recall in a within-experiment comparison.

[7] The small difference between the current experiment and the combined data from Experiments 1 and 2, whereby the latter but not the former reveals a larger free recall vs restudy clustering difference for structured compared to randomised lists (see Supplementary Information), is presumably a subtle consequence of the overall number of items that participants studied and the taxonomic categories employed.

Mediation analyses

To further quantify the possible role of strategy change in TPNL, we conducted three parallel multiple mediator analyses using PROCESS for SPSS (Hayes, 2013) with Bootstrap bias-corrected 95 % CIs generated using 50,000 bootstrap samples. These analyses combined data from the previous experiments. The first planned analysis asked whether the difference between free recall and restudy in final list recall was mediated by semantic clustering (ARC scores), as predicted by the strategy-change account, and/or the number of prior list intrusions. The second and third exploratory analyses asked whether the difference between free recall and list discrimination in final list recall and the difference between list discrimination and restudy in final list recall were mediated by the same factors.

If the effects of different interim tasks on recall are solely due to the degree of semantic clustering that they induce, then we would expect their effects on recall to be entirely mediated by the indirect effect of clustering. If semantic clustering plays some role, we expect a small, but significant, indirect effect. However, if the differences are produced by alternative mechanisms, then we would expect a robust direct effect of interim task (free recall vs restudy or list discrimination), and little or no mediation due to clustering.

The number of prior list intrusions was included as a mediator variable in the analyses primarily because the difference in final list recall between interim tasks appeared to be inversely proportional to the difference in intrusions. That is, the free recall group had the fewest intrusions, followed by the list discrimination and then the restudy group. Differences between interim tasks in final list recall could therefore be mediated by the number of intrusions. Including prior list intrusions in the analysis also has two additional benefits. Firstly, it allows us to compare and contextualise the magnitude of a potential indirect effect of semantic clustering. Secondly, it allows us to evaluate an alternative account of TPNL, the release from proactive interference (PI) theory (Bäuml, & Kliegl, 2013; Szpunar et al., 2008; Yang et al., 2022). The release from PI theory argues that TPNL is caused by testing preventing the build-up of PI, which negatively impacts new learning. As intrusions have been used in the past as a measure of PI (e.g., Szpunar et al., 2008), this account provides a mechanism that could explain an indirect effect of intrusions on the differences between interim tasks.

It is important to note that whereas we are treating intrusions and semantic clustering as causal factors within the mediation model – that is, TPNL is caused by a reduction in intrusions and/or an increase in semantic clustering – it is possible that decreased intrusions and increased semantic clustering are the result of TPNL, or are by-products of TPNL, resulting from some entirely different mechanism (Ahn & Chan, 2022). In addition, whereas intrusions have been used as a measure of PI in the past, they are not a direct measure of PI and the relationship between intrusions and PI is likely to be nonlinear (Postman & Underwood, 1973). This makes interpretation complex. We consider this

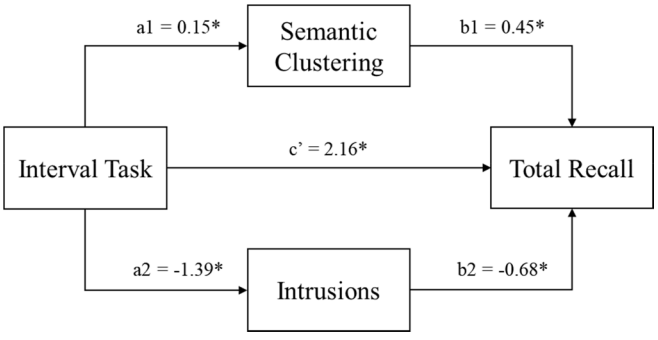

Fig. 5. Semantic clustering and intrusions as mediators for the effect of interim task (Free Recall or Restudy) on total recall in Experiments 1–3. Unstandardised coefficients are shown. * p <.002.

further in the General Discussion. Given these considerations, the mediation analyses should therefore primarily be viewed as examining the extent of the associations between these factors and TPNL, which is an indicator of the strength of the potential relationship.

Free recall vs Restudy on final list recall

We conducted a parallel multiple mediator analysis pooling the free recall and restudy groups from Experiments 1–3, averaging over word order ($N = 494$). See Fig. 5 for a schematic of the mediation model and unstandardised coefficients (B).

The analysis revealed that free recall potentiated new learning relative to restudy indirectly through increasing semantic clustering (a1b1 = 0.07, 95 % CI[0.006, 0.15]) and by reducing prior list intrusions (a2b2 = 0.95, 95 % CI[0.72, 1.20]). A pairwise contrast found that the indirect effect of intrusions was significantly larger than that of semantic clustering (point-difference = 0.88, 95 % CI[0.63, 1.15]). Importantly, there was also evidence that free recall potentiated new learning independently of either semantic clustering or prior list intrusions (c' = 2.16, 95 % CI[1.62, 2.70]). Prior-list free recall tests boosted the number of items remembered by more than 2 items on average, even when statistically controlling for semantic clustering and prior list intrusions.

In other words, this multiple mediator analysis demonstrated that there was an association between both release from PI and conceptual strategy change and TPNL, but that neither association fully accounts for the effect. Furthermore, the association with conceptual strategy change is weaker than that of release from PI, as indicated by the significantly smaller indirect effect of semantic clustering compared to the indirect effect of prior list intrusions. Whereas free recall boosted TPNL by about 1 item relative to restudy via its effect on intrusions, the comparable boost mediated by conceptual strategy change was less than 0.1 items recalled.

Free recall vs list discrimination on final list recall

A second parallel multiple mediator analysis pooled the free recall and list discrimination groups from Experiments 1 and 3, averaging over word order ($N = 412$).

Fig. 6 shows the mediation model and unstandardised coefficients (B). This analysis revealed that the indirect effect of semantic clustering was not significant, a1b1 = 0.17, 95 % CI[-0.026, 0.04]. The indirect effect of prior list intrusions however was significant, a2b2 = 0.13, 95 % CI[0.07, 0.19]. The pairwise contrast was also significant, point-difference = 0.11, 95 % CI[0.05, 0.18]. Hence only release from PI played a significant role in the difference between free recall and list discrimination. Importantly, there was also evidence that compared to list discrimination, free recall produced greater new learning independently of either semantic clustering or intrusions, c' = 1.54, 95 % CI[0.95, 2.13].

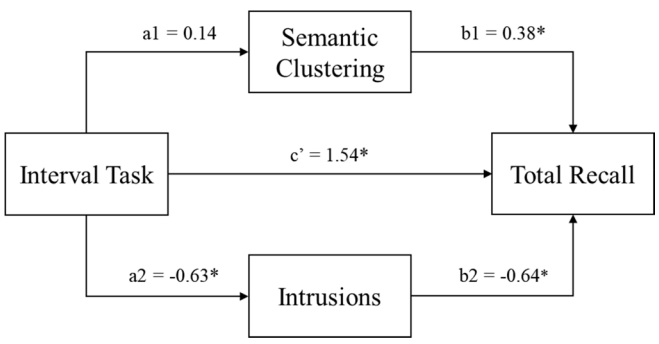

Fig. 6. Semantic clustering and intrusions as mediators for the effect of interim task (Free Recall or List Discrimination) on total recall in Experiments 1 and 3. Unstandardised coefficients are shown. *p <.05.

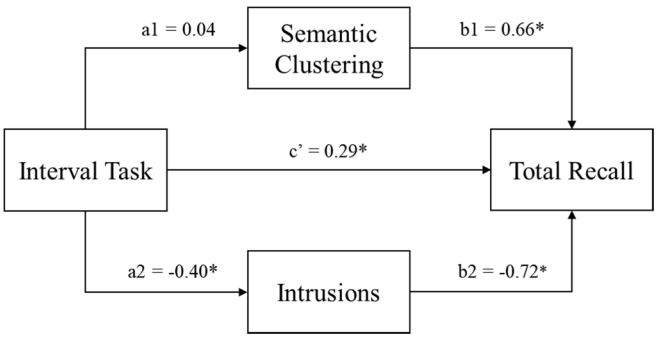

Fig. 7. Semantic clustering and intrusions as mediators for the effect of interim task (List Discrimination or Restudy) on total recall in Experiments 1 and 3. Unstandardised coefficients are shown. *p <.001.

Overall, this analysis found that there was no association between conceptual strategy change and the difference in final recall between groups engaging in interim list discrimination and free recall, but release from PI was associated with this difference. However, release from PI did not fully account for the effect.

List discrimination vs restudy on final list recall

The final analysis pooled the restudy and list discrimination groups from Experiments 1 and 3, averaging over word order ($N = 449$). Fig. 7 shows the mediation model and unstandardised coefficients (B).

Similar to the previous model, the indirect effect of semantic clustering was not significant, a1b1 = -0.03, 95 % CI[-0.08, 0.16] whereas the indirect effect of prior list intrusions was, a2b2 = 0.29, 95 % CI[0.17, 0.42], and the pairwise contrast was significant too, point-difference = 0.31, 95 % CI[0.18, 0.45]. These results indicate that release from PI played a significantly larger role than conceptual strategy usage in the difference in final recall between groups completing interim list discrimination and restudy. When accounting for these indirect effects, the difference between list discrimination and restudy in total recall was still significant, c' = 0.29, 95 % CI[0.04, 0.58].

In summary, this final analysis suggests that whereas release from PI was significantly associated with the difference between list discrimination and restudy, conceptual strategy usage was not. Once accounting for these mechanisms, the difference between the list discrimination and restudy groups persisted.

General Discussion

The current research was conducted to evaluate the degree to which conceptual strategy change contributes to TPNL. The account proposes that one mechanism by which previous retrieval attempts potentiate new learning is by promoting the use of conceptual learning strategies. These strategies make the semantic or taxonomic associations between words and lists more obvious and potentiate learning as new materials are integrated into already existing conceptual categories, increasing both encoding and retrieval efficiency. The results from three experiments suggest that although a conceptual strategy change may be a product of retrieval practice, it is unlikely to be a causal mechanism for TPNL.

We evaluated the strategy change account by testing three of its core predictions. Firstly, a retrieval task (i.e., list discrimination) which does not facilitate the use of conceptual strategies (i.e., does not enhance semantic clustering of recall output) should not potentiate new learning. Secondly, TPNL should be larger when using related compared to unrelated materials, as the number of conceptual associations is increased.

Lastly, TPNL should be reduced when using a list structured around taxonomic categories, as the conceptual associations between words and lists are equally obvious for those in the restudy and retrieval groups.

Table 2
Summary of predictions and support of the conceptual strategy change account.

Prediction	Result
Free recall potentiates new learning and increases the semantic clustering of recall output, relative to restudy	✔
The difference between free recall and restudy will be moderated by the degree of relatedness within and between word lists	×
A retrieval task not promoting semantic clustering (i.e., list discrimination) will not potentiate new learning	×
Structuring lists according to taxonomic categories will attenuate the difference between free recall and restudy in new learning	×
Differences in semantic clustering should align with differences in free recall	×
The difference in final list recall between free recall and restudy will be moderated by semantic clustering	✔

Note. Ticks denote effects consistent with the account, crosses denote inconsistent effects

Table 1 summarises the key qualitative findings from Experiments 1–3, and Table 2 summarises whether the results support the strategy change account.

Although we replicated the primary findings from Chan, Manley et al. (2018), that is, testing did potentiate new learning and resulted in greater semantic clustering of recall output than restudy and list discrimination, the results of our experiments do not provide good evidence of a causal role of strategy change in producing TPNL. Semantic clustering appears to be neither necessary nor sufficient for TPNL. Instead, the increased semantic clustering in free recall conditions likely indicates that increased semantic clustering is a by-product of TPNL.

Experiments 1 and 3 demonstrated that a list discrimination task, which did not increase semantic clustering relative to restudy, also potentiated new learning. List discrimination did not improve recall to the same degree as free recall, which may indicate a potential contribution of strategy change to the effect. However, overall, the evidence suggests this is not the case.

Although research on the testing effect has found that list discrimination tasks decrease overall semantic clustering (and increase temporal clustering) relative to restudy (Whiffen & Karpicke, 2017), our results demonstrate that list discrimination has no impact on semantic clustering in TPNL. That list discrimination tasks potentiate new learning when they do not increase the adoption of conceptual encoding strategies indicates that conceptual strategy change is not necessary for TPNL to occur, and therefore potentiated new learning likely occurs due to different mechanisms, even when word lists are conceptually related. A mediation analysis of the difference in new learning between free recall and list discrimination also highlighted that there was no indirect effect of semantic clustering, suggesting that some other mechanism drives this difference – potentially involving protection from proactive interference.

In all experiments, manipulations that increased semantic clustering did not increase the magnitude of TPNL, indicating that producing greater conceptual strategy change is not sufficient to produce greater TPNL. In Experiment 1, we found that robust TPNL effects of a similar magnitude existed for high-relatedness and mixed-relatedness lists, despite there being substantially greater clustering in recall output for high-relatedness lists. Furthermore, the magnitude of TPNL was similar for highly related, medium related, and unrelated words. This is important as it suggests that the number of possible intra- and inter-list associations did not moderate TPNL. TPNL is of course found when using lists of unrelated words, which is confirmed in a recent meta-analysis (Chan, Meissner, and Davis, 2018). However, we show that the magnitude of the effect does not differ between related and unrelated word lists, contrary to predictions of the strategy change account.

In Experiments 2 and 3 we also found that presenting words in their taxonomic categories did not abolish, nor even reduce, the magnitude of the TPNL effect. We originally hypothesised that structured lists would make the conceptual associations evident to all participants, thereby reducing the difference in semantic clustering between free recall and restudy groups. Instead, structured word lists increased semantic clustering for both groups, and semantic clustering was greater for free recall than restudy conditions. Experiment 3 confirmed that the difference was consistent for randomised and structured word orders. A possible explanation for this could be that although conceptual links between words were obvious for both groups, participants in the restudy group may have paid less overall attention to encoding the lists, which could result in a difference in both new learning and clustering. Alternatively, practicing retrieval may simply enable participants to better use these kinds of strategies.

The clearest example of a dissociation between the degree of conceptual strategy change and new learning is the impact of list structure across Experiments 1 and 2, which was replicated in Experiment 3. Presenting words in their taxonomic structure dramatically increased the amount of semantic clustering in recall output (by nearly 2- and 3-fold respectively) but had no effect on new learning. Additionally, Experiment 3 demonstrated substantially greater semantic clustering of recall output for all groups when using structured word orders, with no discernible enhancement in final list recall. As semantic clustering is used to quantify conceptual strategy usage (Chan, Manley et al., 2018), the fact that all conditions benefitted from substantially greater semantic clustering when using structured word lists, but without any increase in final list recall, suggests that greater adoption of a conceptual strategy conferred no benefit to new learning.

In addition, three mediation analyses also suggested that the association between conceptual strategy change and TPNL was weak. The analysis assessing the difference between free recall and restudy in Experiments 1–3 found that semantic clustering partially mediated the differences between free recall and both interim tasks (i.e., list discrimination and restudy), but this effect was small and TPNL persisted even when accounting for these mechanisms. Furthermore, the indirect effect of semantic clustering was significantly smaller than the indirect effect of prior list intrusions.

The other two analyses on the difference between free recall and list discrimination and list discrimination and restudy in Experiments 1 and 3 found no association between semantic clustering and TPNL, but a large indirect effect of prior list intrusions. Thus the difference in recall between these groups is unlikely due to a causal contribution of conceptual strategy change. Overall, the results from our mediation analyses are consistent with other recent research. Yang et al. (2022) employed unrelated word lists, for which conceptual strategy change should contribute minimally to TPNL, and still observed substantial TPNL. Their analysis also showed a large mediating effect of prior list intrusions. When taken in combination with our experimental manipulations discussed above, these results suggest that the role of conceptual strategy change in TPNL is likely to be either correlational, or minor (or both). It is important to note when interpreting these mediation analyses that they do not establish causality (Ahn & Chan, 2022). It is possible that both semantic clustering and intrusions are simply epiphenomena of testing which are correlated with, but do not cause TPNL. As such, they should be viewed as supplementary to the experimental manipulations employed in this study. Nevertheless, they provide additional support that whatever the precise causal mechanism is, prior list intrusions seem to be more intrinsic to that mechanism than semantic clustering.

Although this study was designed primarily to evaluate the strategy change account of TPNL, the results are also informative regarding the release from PI account of TPNL (Szpunar et al., 2008). In all experiments, alongside robust TPNL, we also observed a substantial and uniform reduction in prior list intrusions. As has been previously argued (Bäuml & Kliegl, 2013; Szpunar et al., 2008; Yang et al., 2022), this reduction in prior list intrusions is evidence that testing prevents the build-up of PI. As the reduction in prior list intrusions was similar across materials, this account is consistent with the uniformity in TPNL magnitude. The release from PI account is also consistent with free recall

producing a larger TPNL effect than list discrimination, as there is a proportionate difference in prior list intrusions. The mediation analyses suggested that a significant portion of the differences in final recall between conditions was associated with reduced PI. However, even when accounting for prior intrusions and semantic clustering, the difference between free recall and the other groups persisted. This suggests that alternative mechanisms must contribute to the effect. However, it should be noted that an exact mechanism regarding how testing reduces PI remains elusive and additional research is needed to clarify this pathway. Recent research has also highlighted that the relationship between intrusions and TPNL is only correlational and that strong evidence from experimental manipulations is lacking (Ahn & Chan, 2022). Additionally, intrusions are at best an indirect measure of PI and the relationship between recorded intrusions and PI is likely nonlinear. For example, previous list words recalled during the criterial test only assess overt intrusions and do not shed light on covert intrusions which participants may consider and reject as incorrect (Postman & Underwood, 1973). Covert intrusions may still accrue proactive interference, but their impact is not quantified. As such, it is unclear whether intrusions are a good measure of the role of PI reduction in TPNL.

There are, of course, multiple alternative mechanisms which are consistent with our results that could explain the TPNL observed, and many of these are discussed in Yang et al.'s (2018) review. For example, testing may potentiate new learning by facilitating the use of other types of strategies. In a recent article, Chan, Manley, and Ahn (2020) argue that testing might potentiate new learning through the optimisation of strategies during encoding and retrieval. Tests improve metacognitive knowledge of what types of questions and content future tests might assess and thus enhance future learning. Testing could increase the use of strategies which allow for more efficient source discrimination, such as facilitating the integration of episodic, temporal cues from the learning environment into a structured memory trace. Such source monitoring, or source discrimination, frameworks have been proposed by Pierce, Gallo, and McCain (2017) and Wahlheim (2015), and a mediation analysis confirms that temporal factors partially mediate TPNL (Yang et al., 2022). These accounts could also explain how list discrimination tasks can potentiate new learning. Additionally, testing might potentiate conceptual strategies that are not measured by semantic clustering. Alternatively, more general mechanisms, such as reducing mind-wandering and increasing attention (Schacter & Szpunar, 2015; Szpunar, Khan, & Schacter, 2013) or the previously discussed context change account could also play a role (Kliegl & Bäuml, 2021).

Of particular note is the recent two-factor account proposed by Kliegl and Bäuml (2021) which proposes that different mechanisms may contribute to TPNL when materials are related and unrelated. Within this framework, conceptual strategy change is the main contributor to TPNL when materials are related, whereas context change is the dominant mechanism when they are not. Our study, which employed lists of related words, implies that even under these conditions, the relative contribution of conceptual strategy change is likely minimal, whereas a potential role of context change is greater – at least when there is a short lag and retention interval. Additionally, it seems likely that the possible role of context change within TPNL is also dependent on the nature of the retrieval task, and the likelihood of semantic processes being engaged. For example, in list discrimination tasks where conceptual strategy change is unlikely, context change might play a larger role than in tasks such as free recall. Additional research is needed to determine whether there is a greater effect of conceptual strategy change when the lag and retention interval are longer, and in exploring alternative tasks which might facilitate even greater conceptual strategy usage.

One final interesting finding from the current study is that these effects were found in a tutorial setting where students were tested on individual devices, but in classrooms of approximately 20, as well as in an online setting. The finding that testing reliably and robustly potentiated new learning in these contexts is demonstrative of the potential pragmatic benefits of retrieval practice in education.

Amongst the limitations of this work, it should be acknowledged that our conclusions about conceptual strategy usage depend on the ARC measure. Although a standard measure (e.g., Yang et al., 2022), it is possible that aspects of conceptual encoding are not adequately captured by this measure. Also, although we employed different list materials across experiments, future work should investigate the generality of our findings with further sets of materials. For example, in Experiment 1, our comparison of related and unrelated lists used different items in the groups (i.e., participants were assessed on learning different words), introducing a possible confound. We emphasize that other key results, including the dissociation between the effects of interim task (list discrimination vs restudy) on final recall and conceptual strategy usage are not compromised by this confound. Nevertheless, future research should attempt to make comparisons where final list items are the same for all conditions. One final note is that our study also employed a short lag between study and test, and a very short retention interval. Future research should look at the potential contribution of a conceptual strategy change when the lag and retention interval are longer.

To conclude, the current study observed consistent and robust TPNL effects using word list materials. The magnitude of TPNL did not appear dependent on whether the word lists were composed of highly-related or mixed-relatedness words. List discrimination tests that did not increase semantic clustering nevertheless induced TPNL. We also observed that free recall tests on learned lists increased the semantic clustering of recall output, and decreased the number of prior list intrusions, relative to restudy. In sum, the results of the current study suggest that conceptual strategy change is not a causal mechanism of TPNL, but rather an effect of retrieval practice.

CRediT authorship contribution statement

Shaun Boustani: Conceptualization, Methodology, Software, Formal analysis, Investigation, Writing – original draft, Writing – review & editing, Supervision, Project administration, Funding acquisition. **Caleb Owens:** Conceptualization, Methodology, Writing – original draft, Supervision. **Hilary J. Don:** Conceptualization, Methodology, Writing – review & editing, Project administration. **Chunliang Yang:** Writing – review & editing, Funding acquisition. **David R. Shanks:** Conceptualization, Methodology, Writing – original draft, Writing – review & editing, Supervision, Project administration, Funding acquisition.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The data and analysis code for this article can be found here: https://osf.io/a25m9/

Acknowledgments

This research was supported by grants from the United Kingdom Economic and Social Research Council (ES/S014616/1) and the Natural Science Foundation of China (32000742).

Appendix A. Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jml.2023.104412.

References

Ahn, D., & Chan, J. C. K. (2022). Does testing enhance new learning because it insulates against proactive interference? *Memory & Cognition, 50*, 1664–1682. https://doi.org/10.3758/s13421-022-01273-7

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

Boustani, S., & Shanks, D. (2022). Heterogeneity and publication bias in research on test-potentiated new learning. *Collabra: Psychology, 8*, 31996. https://doi.org/10.1525/collabra.31996

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

Chan, J., Manley, K., Davis, S., & Szpunar, 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. https://doi.org/10.1016/j.jml.2018.05.007

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

Divis, K. M., & Benjamin, A. S. (2014). Retrieval speeds context fluctuation: Why semantic generation enhances later learning but hinders prior learning. *Memory & Cognition, 42*(7), 1049–1062. https://doi.org/10.3758/s13421-014-0425-y

Faul, F., Erdfelder, E., Buchner, A., & Lang, A. (2009). Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods, 41*(4), 1149–1160. https://doi.org/10.3758/brm.41.4.1149

Hayes, A. (2013). Introduction to mediation, moderation, and conditional process analysis: A regression-based approach. *Guilford*.

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

Jonker, T. R., Seli, P., & MacLeod, C. M. (2013). Putting retrieval-induced forgetting in context: An inhibition-free, context-based account. *Psychological Review, 120*(4), 852–872. https://doi.org/10.1037/a0034246

Kliegl, O., & Bäuml, K. H. T. (2021). When retrieval practice promotes new learning – The critical role of study material. *Journal of Memory and Language, 120*, Article 104253. https://doi.org/10.1016/j.jml.2021.104253

Kliegl, O., Kriechbaum, V. M., & Bäuml, K.-H.-T. (2022). The effects of interspersed retrieval practice in multiple-list learning on initially studied material. *Frontiers in Psychology, 13*. https://doi.org/10.3389/fpsyg.2022.889622

Lau, M. C., Goh, W. D., & Yap, M. J. (2018). An item-level analysis of lexical-semantic effects in free recall and recognition memory using the megastudy approach. *Quarterly Journal of Experimental Psychology (2006), 71*(10), 2207–2222. https://doi.org/10.1177/1747021817739834

Nunes, L. D., & Weinstein, Y. (2012). Testing improves true recall and protects against the build-up of proactive interference without increasing false recall. *Memory, 20*(2), 138–154. https://doi.org/10.1080/09658211.2011.648198

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

Pastötter, B., Bäuml, K. H., & Hanslmayr, S. (2008). Oscillatory brain activity before and after an internal context change — Evidence for a reset of encoding processes. *NeuroImage, 43*(1), 173–181. https://doi.org/10.1016/j.neuroimage.2008.07.005

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

Pierce, B., Gallo, D., & McCain, J. (2017). Reduced interference from memory testing: A postretrieval monitoring account. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 43*(7), 1063–1072. https://doi.org/10.1037/xlm0000377

Postman, L., & Underwood, B. J. (1973). Critical issues in interference theory. *Memory & Cognition, 1*(1), 19–40. https://doi.org/10.3758/BF03198064

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.

Schacter, D., & Szpunar, K. (2015). Enhancing attention and memory during video-recorded lectures. *Scholarship of Teaching and Learning in Psychology, 1*(1), 60–71. https://doi.org/10.1037/stl0000011

Sahakyan, L., & Kelley, C. M. (2002). A contextual change account of the directed forgetting effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 28*(6), 1064–1072. https://doi.org/10.1037/0278-7393.28.6.1064

Szpunar, K., Khan, N., & Schacter, D. (2013). Interpolated memory tests reduce mind wandering and improve learning of online lectures. *Proceedings of the National Academy of Sciences of the United States, 110*(16), 6313–6317. https://doi.org/10.1073/pnas.1221764110

Szpunar, K., McDermott, K., & Roediger, H. (2008). Testing during study insulates against the buildup of proactive interference. *Journal of Experimental Psychology: Learning, Memory and Cognition, 34*(6), 1392–1399. https://doi.org/10.1037/a0013082

Tulving, E. (1962). Subjective organization in free recall of "unrelated" words. *Psychological Review, 69*(4), 344–354.

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

Whiffen, J., & Karpicke, J. (2017). The role of episodic context in retrieval practice effects. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 43*(7), 1036–1046. https://doi.org/10.1037/xlm0000379

Wissman, K., Rawson, K., & Pyc, M. (2011). The interim test effect: Testing prior material can facilitate the learning of new material. *Psychonomic Bulletin & Review, 18*(6), 1140–1147. https://doi.org/10.3758/s13423-011-0140-7

Van Overschelde, J., Rawson, K., & Dunlosky, J. (2004). Category norms: An updated and expanded version of the Battig and Montague (1969) norms. *Journal of Memory and Language, 50*(3), 289–335. https://doi.org/10.1016/j.jml.2003.10.003

Yang, C., Luo, L., Sun, B., Zhao, W., Potts, R., & Shanks, D. R. (2022). Testing potential mechanisms underlying test-potentiated new learning. *Journal of Experimental Psychology. Learning, Memory, and Cognition, 48*, 1127–1143. https://doi.org/10.1037/xlm0001021

Yang, C., Potts, R., & Shanks, D. (2018). Enhancing learning and retrieval of new information: A review of the forward testing effect. *NPJ Science of Learning, 3*(1). https://doi.org/10.1038/s41539-018-0024-y, 8–8.