Do partial and distributed tests enhance new learning?

Hilary J. Don[1], Chunliang Yang[2], Shaun Boustani[1] & David R. Shanks[1]

[1]Division of Psychology and Language Sciences, University College London

[2]Institute of Developmental Psychology, Beijing Normal University

Correspondence should be addressed to Hilary J. Don
Division of Psychology and Language Sciences
University College London
26 Bedford Way, Bloomsbury, London WC1H 0AP, United Kingdom
h.don@ucl.ac.uk

Abstract

Testing facilitates subsequent learning of new information, a phenomenon known as the *forward testing effect*. The effect is often investigated in multi-list procedures, where studied lists are followed by a retrieval test, or a control task such as restudying, and learning is compared on the final list. In most studies of the effect, tests include all material from the preceding list. We report four experiments, three of which were pre-registered, to determine whether tests that are partial (not including all studied items) and distributed (including retrieval of items from earlier lists) are effective in enhancing new learning. The results show that testing of all studied material is not necessary to produce beneficial effects on new learning, or to reduce intrusions. The beneficial effects of testing were substantially mediated by reduced proactive interference. Importantly, there was minimal evidence that the forward learning benefits of partial and distributed tests are offset by a cost to untested items via retrieval-induced forgetting.

Hundreds of laboratory and classroom studies have established that testing can more effectively enhance learning than other study strategies, such as restudying, a phenomenon referred to as *test-enhanced learning* (for reviews, see Rowland, 2014; Yang et al., 2021). For instance, retrieval practice has been shown to facilitate consolidation and long-term retention of studied information when compared to other strategies, the *backward testing effect* (BTE; see Roediger & Karpicke, 2006, for a review). More recently it has been discovered that testing also facilitates learning and retention of new information, the *forward testing effect*[1] (FTE; Chan et al., 2018; Szpunar, McDermott, & Roediger, 2008; Yang et al., 2018).

The FTE has been repeatedly demonstrated in multi-list procedures. For instance, participants in experiments by Weinstein et al. (2011) and Yang et al. (2017) studied four lists of face–name pairs. After studying each of lists 1-3, a test group practised retrieval of the pairs in the previous list whereas a control (no-test) group solved math problems. In the interim tests, all faces from the just-studied list were shown one-by-one and participants were asked to recall their corresponding names. After list 4, both groups were asked to recall target names from that list. Participants in the test group correctly recalled more targets than the no-test group, indicating that prior testing enhanced new learning of list 4. The effect has been reliably replicated using a variety of study materials using both free- and cued-recall, including foreign language word pairs (Cho, Neely, Crocco, & Vitrano, 2016; Yang et al., 2017; 2019), word lists (Szpunar et al., 2008), text passages (Wissman, Rawson, & Pyc, 2011), and video lectures (Jing, Szpunar & Schacter, 2016; Schacter & Szpunar, 2015; Szpunar, Khan, & Schacter, 2013). It is also broadly present in many different populations, including young children (Aslan & Bäuml, 2016), college students (e.g., Szpunar et al., 2008), older adults (Wang, Yang & Zhong, 2020), and traumatic

[1] Also termed *test-enhanced* or *test-potentiated new learning* (Chan et al., 2018).

brain injury patient groups (Pastötter, Weber, & Bäuml, 2013). The effect therefore has broad applications in educational practice.

Several theories have been proposed to explain the effect of testing on new learning (for reviews, see Chan et al., 2018, and Yang et al., 2018). According to the *release from proactive interference* (PI) theory, interim tests induce context changes which reduce the build-up of PI (Szpunar et al., 2008). Tested words become associated with both a study context and a retrieval context, while newly studied words are only associated with a study context. There will therefore be greater differentiation between new and previously studied pairs, which facilitates recall of newly learned materials (Karpicke et al., 2014, Szpunar et al., 2008). Similarly, *reset of encoding* theories also propose that interim tests induce context changes that "reset" encoding between learning blocks, such that future encoding can occur just as effectively as previous encoding (Pastötter et al., 2011).

The number of intrusions (incorrectly recalling words from prior lists) is taken as an index of PI. Several studies provide support for the release from PI explanation, showing that the FTE occurs concurrently with reduced intrusions (e.g., Aslan & Bäuml, 2016; Bufe & Aslan, 2018; Nunes & Weinstein, 2012; Pastötter & Bäuml, 2014; Pastötter et al., 2013; Pierce et al., 2017; Szpunar et al., 2008l Weinstein et al., 2011; Weinstein et al., 2014; Yang et al., 2017; Yang et al., 2019). Moreover, Yang et al. (2022) demonstrated that the effect of the interim task on correct recall was mediated at least in part by prior list intrusions, indicating that release from PI contributes to the effect.

In contrast, strategy change theories propose that repeated experience with tests allow participants to develop and use more effective encoding (Cho et al., 2017) or retrieval strategies (Soderstrom & Bjork, 2014). In addition to their evidence on release from PI, Yang et al. (2022)

found that the effect of the interim task on correct recall was partly mediated by such strategy changes. Other theories suggest that interim testing increases expectancy of receiving a test, which motivates study effort for new material, or enhanced attentional encoding of new information (Weinstein et al., 2014). These mechanisms may be non-exclusive, such that the forward testing effect is a result of multiple processes.

In the standard FTE design adopted in most experiments conducted to date, the interim tests have tested all pairs from the preceding list. In the Weinstein et al. (2011) and Yang et al. (2017) experiments described above, for example, each list comprised 12 face-name pairs, all of which (for the test group) were tested via cued recall in the interim test. List learning experiments employing free rather than cued recall in the interim tests (e.g., Szpunar et al., 2008) also ask participants to recall all targets from the previous list. However, it is unlikely in classroom settings that tests will cover all information studied in the immediately-preceding lesson. Yue et al. (2015, Experiment 2) reported an experiment addressing the important question of whether partial as opposed to full tests are sufficient to induce forward testing effects. In their experiment, Yue et al. presented participants with a short animated and narrated video describing the life cycle of a star. One group was then tested on half the key concepts in the video via cued recall questions, while another group restudied these concepts. All participants then watched a second video about lightning formation, and finally took a test on the content of this second video. Yue et al. found a forward testing effect: Participants who had been partially tested on the first video learned the second video more effectively, answering more questions correctly than a group that had not been tested.

Yue et al.’s (2015) experiment provides a starting point to address whether partial tests effectively induce a forward testing effect. In this study, we aimed to extend this research and address three novel questions that remain unanswered by Yue et al. First. while Yue et al. demonstrated enhanced learning after partial testing, this was not compared to a full-test condition, such that we cannot determine whether partial tests enhance learning to the same extent as a full test. Second, in educational settings, tests will not only include only some of the learned material, but will also include material taught over multiple separate classes and lessons. The present study therefore aimed to test whether distributed tests (when a test requires retrieval from earlier lists, not just the immediately preceding one) are also effective in potentiating new learning. If distributed tests do not potentiate new learning, the practical applications of the forward testing effect will be limited. Yue et al. could not look at distribution because their design only included one learning phase prior to the target phase. Third, the present experiments ask whether *retrieval-induced forgetting* (see Bäuml & Kliegl, 2017, for a review) offsets any tendency for partial tests to potentiate subsequent learning. retrieval-induced forgetting refers to the phenomenon in which retrieval practice of a subset of material impairs subsequent recall for untested pairs. By its very design, any benefits of partial testing may come at a cost to memory for untested pairs. Because they used semantically related materials for which retrieval-induced forgetting is unlikely to occur (see Chan, 2009), Yue et al.'s experiment does not address this potential limitation. We therefore used different materials – but still of educational relevance – namely face-name associations and foreign language vocabulary[2].

The current experiments did not set out to directly test different theories of the forward testing effect, but they do provide an opportunity to examine whether release from PI (Szpunar et

[2] Experiments by Cho et al. (2017, Experiment 3) and Bolte (2019) included conditions in which (as in our experiments) participants studied paired-associates, were tested on some items, and then learned new pairs. However, because their research question was different, they did not include a control condition to assess whether the partial test enhanced learning of the final list.

al., 2008) is influenced by partial testing.[3] There are two potential predictions we can make about the effects of partial and distributed test groups on PI. On the one hand, partial and distributed lists could reduce context differentiation between old and new lists, as untested pairs will only be associated with a study context, and some tested pairs may be associated with multiple test contexts. One might assume that such conditions may be less effective in reducing PI. On the other hand, Pastötter et al. (2011) found that simply generating exemplars from an unrelated category can reduce PI, suggesting that retrieval from long-term memory, even if not from the just-studied list, can increase list discrimination. It therefore may be reasonable to expect that partial testing is sufficient to reduce PI. Here, we conduct a mediation analysis similar to Yang et al. (2022) to determine whether prior list intrusions mediate the effect of testing on recall in each test condition. This type of analysis has not yet been conducted to test the influence of release from PI on the FTE with paired-associate material.

Experiment 1

Experiment 1 aimed to test whether partial tests can enhance new learning in a multi-list design with face-name pairs (Weinstein et al., 2011; Yang et al., 2017). Participants learned four sets of 12 face-name pairs. Two groups replicate a typical FTE task, in which a Full-Test group received a test on all studied face-name pairs, and a Restudy group reviewed all face-name pairs

[3] The theoretical frameworks of the forward testing effect mentioned here do not make clear predictions about the magnitude of the FTE following partial tests, and each framework could provide an explanation for results in either direction. That is, partial tests may or may not be sufficient to induce release from PI, reset of encoding, strategy changes, or expectancy changes that may produce the FTE. We focus on the release-from-PI hypothesis specifically as the designs of the experiments allow an index of PI, such that we can assess whether it mediates group differences. Expectancy ratings are included in Experiment 1 (see Supplemental Material), however we do not have measures to index strategy change or reset of encoding, and therefore cannot test hypotheses about the effect of partial tests on these phenomena.

after studying each list. In a third Partial-Test group, half the pairs were selected to be tested after each list. Each of the six tested pairs was presented twice in the interim test, to match the number of test trials in the Full-Test group. The design is illustrated in Figure 1. We expected to replicate the FTE, where criterial test recall is greater in the Full-Test than Restudy group. If partial tests are also sufficient to enhance new learning, we should also expect improved recall in the Partial-Test group compared to the Restudy group.

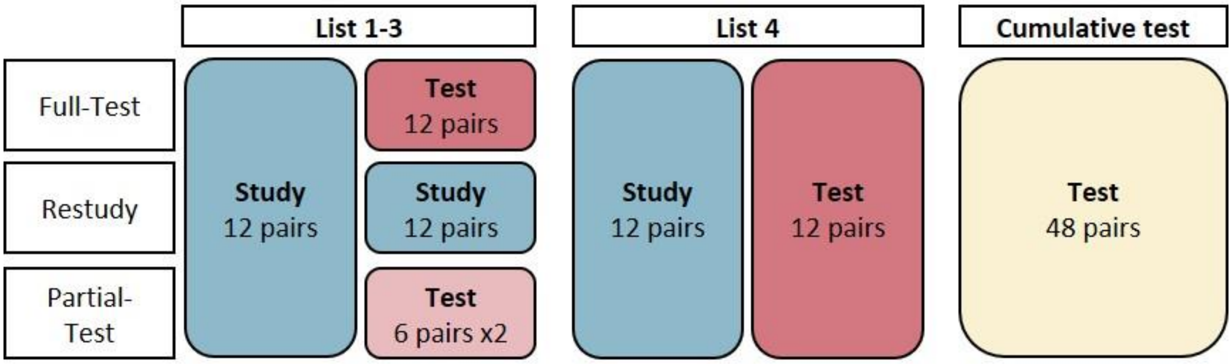

Figure 1. Schematic of the design of Experiment 1.

Method

Participants

We calculated the required sample size to achieve a 95% confidence interval no wider than ±0.45 on an estimated medium effect size of $d = 0.5$. A sample size of 40 participants per group was determined. One hundred and twenty participants were recruited via Prolific (www.prolific.co). The data from two participants were excluded for indicating they had taken notes or recordings during the task, and the data from one participant were excluded for making no serious attempt at recall in the cumulative test. The final sample included 117 participants (81 female, mean age = 36.03, *SD* = 13.76), with 40 participants in the Test group, 38 in the Partial test group, and 39 in the Restudy group.

Materials

All experiments reported here were programmed in PsychoPy and run online via Qualtrics and Pavlovia, and were approved by the Ethics Committee of the UCL Department of Experimental Psychology (EP/2020/007). Participants provided consent for taking part. Materials included 4 sets of 12 face-name pairs. Faces were sourced from the Face Research Lab London Set (DeBruine & Jones, 2017). Names were drawn from a list of the 100 most popular male and female names in the United Kingdom in 2019. Faces were allocated to sets so that each set had six male and six female faces, with diverse ethnicities. Names (all unique) were divided into sets and randomly allocated to faces. Each set of face-name pairs was randomly allocated to list order prior to commencing the study, and this random list order was used for all participants. Within each list, face-name pairs were presented in random order.

Procedure

Participants were instructed that they would learn lists of face-name pairs. Prior to each list, they were asked to rate their expectancy that they would restudy or receive a test for the following list, on a scale from 0 (*definitely review*) to 10 (*definitely test*). On each trial, each face image was presented in the center of the screen, with the corresponding name below it. Study was self-paced in order to be more applicable to real-world settings, as well as allowing an index of study effort between groups. Participants studied the pair before pressing a button to continue to the next pair. If participants took longer than 30 seconds on any given trial, the program automatically progressed to the next trial. Participants were instructed that after studying each list, the computer would randomly decide whether they would receive a short test or a second opportunity to learn the list. In fact, for the first three lists, participants in the Full-Test and Partial-Test group always received a test, and participants in the Restudy group always reviewed the list. For each test, face images were presented one-by-one in random order, and participants were prompted to type in the corresponding name and press enter to move to the next trial. There was no time limit for test trials. In the Restudy group, participants were able to review each face-name pair again, for up to 30 seconds each.

In the Full-Test group, the interim tests included all 12 pairs from the preceding list. In the Partial test group, 6 pairs from each list were selected to be tested. Each pair was then presented twice in random order to match the number of test trials in the Test group. After studying the fourth list, all participants regardless of group received a test including all 12 pairs from List 4. All participants were then administered a cumulative test, comprising all 48 face-name pairs in random order. There was no feedback in the interim and cumulative tests. They were then asked to indicate whether they had taken any notes or photos to assist their performance during the study.

Transparency and Openness

For all experiments, we report how we determined sample size, data exclusions, all manipulations, and all measures in the study. Data were analysed using R version 4.0.5 and JASP version 0.13.1. All summarized data have been made publicly available at the Open Science Framework (OSF) and can be accessed at https://osf.io/3pkj4/. Materials are available in the Supplemental Material. Analysis code is not available, but analyses are described in full in the results section. Experiment 1 was not preregistered.

Data analysis

Correct cue-target recall was considered correct recall. Responses were scored using the *stringdist* package in R. Responses with a maximum string distance of 1 from the correct translation were accepted as correct, responses with a string distance of 2 were screened for typographic errors and scored manually, and responses with a string distance of 3 or more were scored as incorrect.

Intrusions were scored as incorrect recall of names from Lists 1-3 in the criterial test. Response latency can also be used as a measure of PI, where shorter response times should reflect a smaller search set, and therefore less interference (Bauml & Kliegl, 2013; Lehman, Smith & Karpicke, 2014; Wixted & Rohrer, 1993). However, there are several reasons why response times might differ between groups, which are unrelated to the size of the search set. For instance, given that testing enhances learning, shorter response times may be an artefact of better learning, rather than reflecting a mechanism underlying that learning. Response latencies are therefore reported in the Supplemental Material.

We report *p*-values as well as Bayes factors to assess the strength of evidence for the alternative hypothesis (BF_{10}) or null hypothesis (BF_{01}). Bayes factors were computed in JASP using Bayesian analyses of variance (ANOVAs) or *t*-tests with default priors. Typically, *BF*s between 1 and 3 indicate minimal support for the alternative hypothesis, *BF*s between 3 and 10 indicate moderate support, and *BF*s greater than 10 indicate strong support. However, they can also be interpreted continuously as the odds in favour of the alternative hypothesis (Wagenmakers et al., 2018). For Bayesian ANOVA, Bayes factors for the main effects indicate the likelihood of the data given the main effects model relative to a null model (BF_{10}).

Analyses reported in the main text focus primarily on the criterial and cumulative tests. Additional analyses (including study time, interim test performance, expectancy ratings, and additional cumulative test data) are reported in the Supplemental Material and are in line with previous experiments (e.g., Yang et al., 2017). In the following analyses, we first tested for an overall group difference in each measure with a one-way ANOVA, followed by pair-wise comparisons via independent samples *t*-tests. One-tailed *t*-tests were conducted comparing the test groups to the restudy group, as we expected a directional effect of enhanced learning following testing compared to restudying. These compared the Full-Test to Restudy group, to determine whether the FTE is replicated, and the Partial-Test group to the Restudy group, to determine whether partial tests are sufficient to enhance learning. We also compared the Full-Test and Partial-Test groups with a two-tailed independent samples t-test, to determine whether the conditions enhanced learning to the same degree. Holm-Bonferroni adjustments were used to control for multiple comparisons (described *p*-values have been adjusted). Data from all experiments, not including analysis code, are available at the Open Science Framework (OSF; https://osf.io/3pkj4/?view_only=678a6b8440f74194b9a206c3265660bd). Study materials from each experiment are listed in the Supplemental Material.

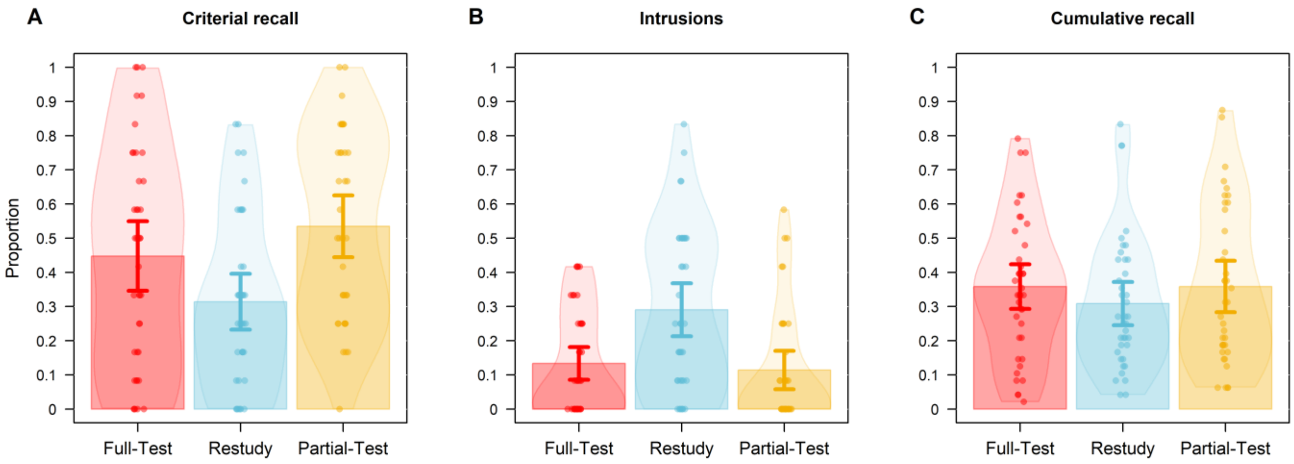

Figure 2. Results from Experiment 1. A: proportion of correct recall in the criterial test. B: proportion of intrusions out of all 12 recall attempts in the criterial test. C: proportion of correct recall in the cumulative test. Error bars represent 95% confidence intervals.

Results & Discussion

Criterial test

Figure 2A shows the proportion of correctly recalled pairs in the criterial test in each group. A one-way ANOVA indicated a significant main effect of Interim Task, $F(2,114) = 5.95$, $p = .003$, $\eta^2_p = .094$, $BF_{10} = 10.02$. One-tailed pairwise comparisons showed a significant standard forward testing effect, with greater recall in the Full-Test group ($M = .45$, $SD =.32$) compared to the Restudy group ($M = .31$, $SD = .25$), $t(77) = 2.07$, $p = .042$, $d = 0.47$, $BF_{10} = 2.84$, replicating earlier studies (e.g., Weinstein et al., 2011; Yang et al., 2017). A novel finding was that partial tests also provided an effective way of enhancing new learning compared to restudying in paired-associate material, with greater recall in the Partial-Test group ($M = .54$, $SD = .28$) than the Restudy group, $t(75) = 3.68$, $p < .001$, $d = 0.84$, $BF_{10} = 122.94$. A two-tailed *t*-test indicated that there was no significant difference in recall in the Full- and Partial-Test groups, $t(76) = 1.29$, $p = .201$, $d = 0.29$, $BF_{01} = 2.07$.

There was also a significant main effect of Interim Task on the proportion of prior-list intrusions in the criterial test (see Figure 2B), $F(2,114) = 10.11$, $p < .001$, $\eta^2_p = .151$, $BF_{10} = 261.28$. One-tailed *t*-tests indicated that testing reduced intrusions compared to restudying for both the Full-Test, $t(77) = 3.52$, $p < .001$, $d = .79$, $BF_{10} = 79.22$, and Partial-Test groups, $t(75) = 3.73$, $p < .001$, $d = 0.85$, $BF_{10} = 141.18$. This suggests that retrieval practice, even if incomplete, can prevent the build-up of PI. There was no significant difference in intrusions between the Full-Test and Partial-Test groups, $t(76) = 0.53$, $p = .596$, $d = 0.12$, $BF_{01} = 3.76$.

Cumulative test

Overall, the benefit of retrieval practice on new learning did not persist into the cumulative test (Figure 2C). There was no significant main effect of Interim Task, $F(2,114)$, $p = .484$, $\eta^2_p = .013$, $BF_{01} = 6.79$. Pairwise comparisons using one-tailed *t*-tests also showed no significant difference between the Full-Test and Restudy groups, $t(77) = 1.10$, $p = .137$, $d = 0.25$,

BF_{01} = 1.49, or Partial-Test and Restudy groups, *t*(75) = 1.03, *p* = .154, *d* = 0.23, BF_{01} = 1.62. A two-tailed *t*-test found no significant difference between the Full-Test and Partial-Test groups, *t*(76) = 0.004, *p* = .996, *d* = 0.001, BF_{01} = 4.26.

To assess retrieval induced forgetting, a mixed effects logistic regression was conducted on the Partial-Test group to analyse cumulative test recall, with participant and pair as random factors, and whether or not pairs were tested as a fixed factor. Face-name pairs were more likely to be recalled correctly if they had been included in interim tests than if they had not, *B* = 0.96, *SE* = 0.21, *z* = 4.56, *p* < .001, *OR* = 2.60, 95% *CI* [1.72, 4.0]. However, a two-tailed *t*-test indicated there was no significant difference in recall between untested pairs in the Partial-Test group (*M* = 0.26, *SD* = 0.21), and untested pairs in the Restudy group (*M* = 0.33, *SD* = 0.21), *t*(75) = 1.49, *p* = .142, *d* = 0.34, BF_{01} = 1.64, although the effect is in the direction that is consistent with retrieval-induced forgetting. We will return to this point later with a full discussion of retrieval-induced forgetting. Additional analyses for cumulative test recall are reported in the Supplemental Material.

Experiment 2

Experiment 1 showed that partial tests that only included some pairs from the just-studied list were as effective as full tests in enhancing new learning. In Experiment 2, we sought to generalize this finding to foreign-language vocabulary learning, and more importantly we aimed to test whether a distributed partial test is also effective in enhancing learning. In educational settings, quizzes frequently cover test questions on not only just-studied, but also previously studied material (e.g., from an earlier class in the semester). Here, the Partial-Test group was replaced with a Distributed-Test group, in which only 8 pairs were tested in each of the first 3 interim tests (here ‘Distributed’ in the group label is shorthand for both distributed and partial).

These tested pairs included some pairs from the list immediately prior, as well as some pairs from preceding lists. The design is shown in Figure 3. If distributed partial tests also benefit new learning, we would expect greater recall of the criterial test pairs in the Distributed-Test group compared to the restudy group, similarly to the Full-Test group.

Experiment 2 also used a delayed cumulative test, administered 24 hours after completing the criterial test, to ascertain whether the benefits of a distributed partial test are maintained after a delay. Chan (2009) demonstrated that retrieval induced facilitation, rather than forgetting, occurred when the final test occurred after a 24-hour delay. The benefit of testing on final test performance also tends to be greater after a delay (Roediger & Karpicke (2006).

Method

Transparency and openness

The experiment's design, hypotheses and analysis plan were pre-registered at OSF, available at osf.io/ayjwd.

Participants

Experiment 2 used a convenience sample of UCL undergraduate students, who participated as part of a laboratory practical class. The maximum sample size was therefore limited to the number of students enrolled in the course, which was approximately 250 participants. In Session 1, we were able to collect data from 181 participants, somewhat fewer than the expected sample. Participants were allocated to groups consecutively in the order they started the study. Data were excluded according to the pre-registered plan if they were incomplete[4] (5 participants), if there were no serious attempts at recall in the criterial test (2 participants), if mean study time in the final list was < 500 ms (1 participant), if participants indicated that they took notes or other recordings during either session (2 participants), or if they indicated prior knowledge of the Euskara language (3 participants).

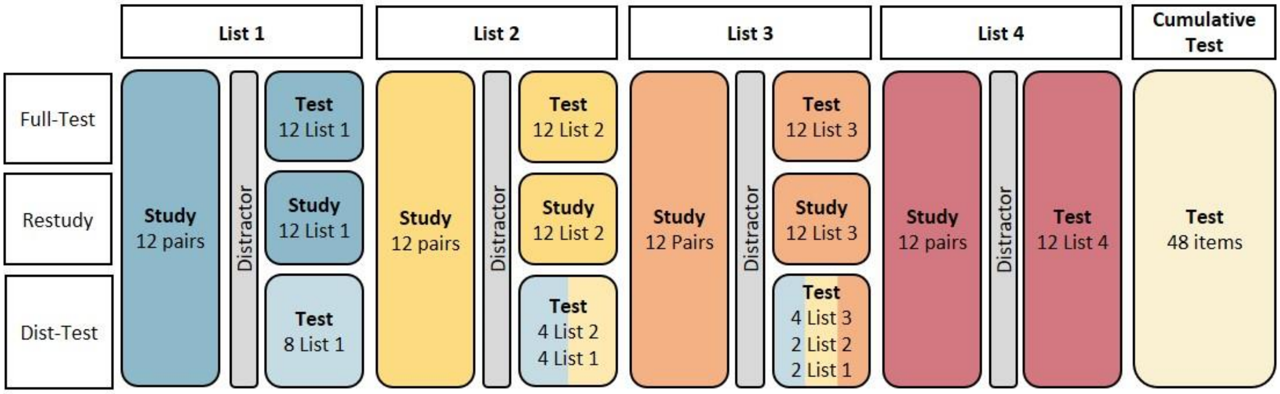

Figure 3. Schematic of the design of Experiment 2.

The final sample for session 1 included 168 participants (mean age = 18.81, *SD* = 1.07, 143 females), with 55 in the Test group, 59 in the Distributed Test group, and 54 in the Restudy group. Nineteen participants completed Session 1 but not Session 2 (5 in the Full-Test group, 7 in the Distributed-Test group, and 7 in the Restudy group). In Session 2, we collected data from 151 participants, and two participants were excluded as they did not have corresponding data from Session 1. This left 50 participants in the Test group, 52 in the Distributed Test group and 47 in the Restudy group. A *post hoc* power analysis using G*Power (Faul, Erdfelder, Lang, & Buchner, 2007) indicated that this sample size would provide 53% power to detect an effect size of *d* = 0.33, which was established in pilot experiments using the same materials. We note that while this is not a high level of power, it was sufficient to observe the primary effects, which are replicated in a higher-powered study in Experiment 3.

[4] This refers to incomplete data from each session. Participants who completed Session 1 but not Session 2 were included in Session 1 analyses.

Materials

Study stimuli included four sets of 12 Euskara-English word pairs. Sets were matched for mean Euskara word length and number of syllables. The allocation of word sets to lists was randomised for each participant, and word pairs within each list were presented in random order.

Procedure

Session 1 was made available to participants between 8am and midnight on the day of testing. Participants were instructed that they would learn lists of Euskara-English word translations. A word pair was presented in the center of the screen on each trial, participants studied the pair before pressing a button to continue to the next pair, and the program automatically progressed after 30 seconds. We also included a short filler task immediately after studying each list, which required participants to evaluate as many true or false arithmetic equations as they could for one minute, before completing the interim task.

In the Full-Test group, the interim tests comprised all 12 pairs from the preceding list. In the Distributed-Test group, the interim test comprised some pairs from the immediately-preceding list, as well as (except for the List 1 test) some pairs from earlier lists, with a total of eight tested pairs. The first test comprised eight pairs from List 1, the second test comprised four pairs from List 1 and four from List 2, and the third test comprised four pairs from List 3, two from List 2, and two from List 1. Each tested pair was drawn randomly from the relevant previous lists. In each test, the Euskara word was presented on screen and participants were prompted to type the English translation, before pressing enter to continue to the next test trial. As in Experiment 1, there was no time limit for test trials. The Restudy group had a second opportunity to study all pairs from the previous list.

After studying the fourth list, all participants regardless of group received a test including all 12 pairs from List 4. Participants were then asked to indicate whether they had taken any notes or recordings during the study, and if they had any prior knowledge of Euskara before participating in the study. They were then reminded to complete Session 2 in 24 hours. Session 2 was made available to participants between 8am and midnight the following day, and participants were requested to complete Session 2 as close to 24 hours after Session 1 as possible The mean delay was 25.10 hours (*SD* = 2.8 hours, *range* = 18.0 - 39.6) All participants completed a cumulative test of all 48 pairs in random order. After completing the cumulative test, participants were again asked to indicate whether they had taken any notes or recordings to assist their performance during the study.

Results & Discussion

The study and analyses proceeded in accordance with the pre-registration plan. Analyses for study time and interim tests are reported in the Supplemental Material.

Criterial Test

Experiment 2 replicated the results from Experiment 1 with a partial and distributed test group. There was a significant main effect of Interim Task, *F*(2,165) = 3.73, *p* = .026, η^2_p = .043, BF_{10} = 1.45, and planned comparisons with one-tailed *t*-tests indicated significantly greater recall in the Full-Test group (*M* = .63, *SD* = .29) than the Restudy group (*M* = .47, *SD* = .30), *t*(107) = 2.77, *p* = .003, *d* = .53, BF_{10} = 11.60, as well as significantly greater recall in the Distributed-Test *(M* = .59, *SD* = .33) than the Restudy group, *t*(111) = 1.89, *p* = .031, *d* = 0.36, BF_{10} = 1.90 (see Figure 4A). We note however that the *BF*s for all but one of these effects indicate weak evidence. A two-tailed *t*-test found that there was no significant difference in criterial test recall in the Full-Test and Distributed-Test groups, $t(112) = 0.74$, $p = .464$, $d = .138$, $BF_{01} = 3.94$[5].

A similar pattern was also observed for prior list intrusions as seen in Experiment 1 (see Figure 4B). There was a significant main effect of Interim Task, $F(2,165) = 9.66$, $p < .001$, $\eta^2_p = .11$, $BF_{10} = 204.65$. Pairwise comparisons with one-tailed *t*-tests showed significantly fewer intrusions in the Full-Test compared to the Restudy group, $t(107) = 3.67$, $p < .001$, $d = 0.70$, $BF_{10} = 133.96$, and significantly fewer intrusions in the Distributed-Test compared to the Restudy group, $t(111) = 2.92$, $p = .002$, $d = 0.55$, $BF_{10} = 16.71$. A two-tailed *t*-test showed no significant difference in intrusions between the Full-Test and Distributed-Test groups, $t(112) = 1.33$, $p = .187$, $d = 0.25$, $BF_{01} = 2.27$.

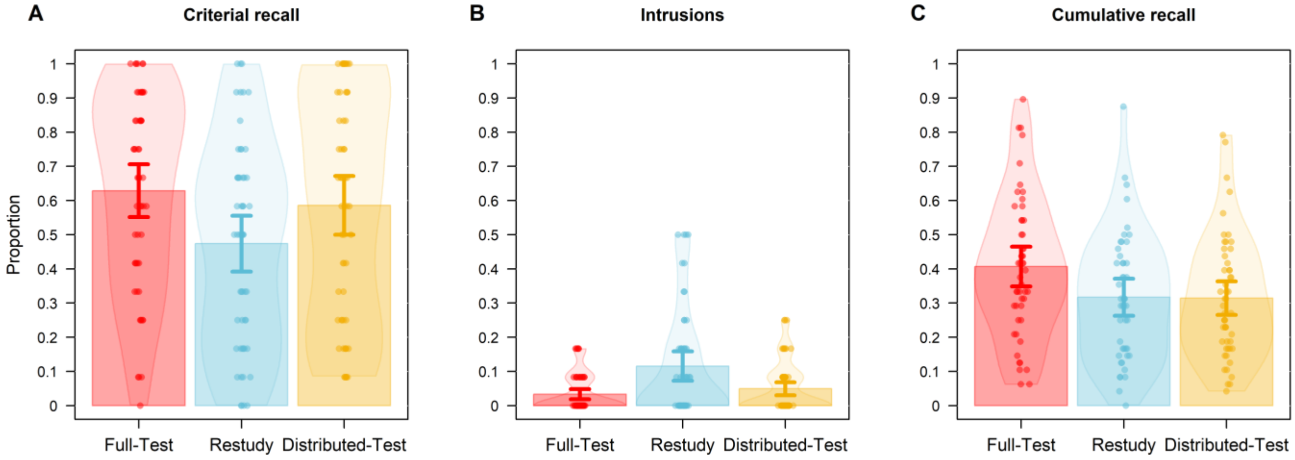

Figure 4. Results from Experiment 2. A: proportion of correct recall in the criterial test. B: proportion of intrusions in the criterial test. C: proportion of correct recall in the cumulative test. Error bars represent 95% confidence intervals.

Cumulative test

[5] As the two one-tailed t-tests were planned comparisons included in the preregistration, we did not use Holm-Bonferroni corrections in Experiment 2.

In the cumulative test, there was a significant main effect of Interim Task, $F(2,146) = 3.85$, $p = .023$, $\eta^2_p = .05$, $BF_{10} = 1.72$. One-tailed pairwise comparisons indicated greater recall in the Full-Test group compared to the Restudy group, $t(95) = 2.26$, $p = .013$, $d = 0.46$, $BF_{10} = 3.93$, but no significant difference in recall between the Distributed-Test and Restudy groups, $t(97) = 0.07$, $p = .527$, $d = 0.01$, $BF_{10} = 0.20$ (see Figure 4C). Thus, the benefit in recall was maintained after a 24 hour delay for the Full-Test group, but not for the Distributed-Test group. A two-tailed *t*-test indicated significantly better recall in the Full-Test group compared to the Distributed-Test group, $t(100) = 2.44$, $p = .017$, $d = 0.48$, $BF_{10} = 2.82$.

To determine how interim testing influenced cumulative test recall in the Distributed-Test group, we ran a mixed effects logistic regression to predict correct recall, with number of times pairs were tested as a fixed factor, and participant and pair as random factors. This showed that pairs that were tested more frequently in the interim tests were more likely to be correctly recalled in the cumulative test, $B = 1.12$, $SE = 0.09$, $z = 12.89$, $p < .001$, $OR = 3.05$, 95% *CI* [2.58, 3.64]. However, testing did not benefit untested pairs. On the contrary, a two-tailed *t*-test indicated correct recall for untested pairs was lower in the Distributed-Test group ($M = 0.17$, $SD = 0.14$) compared to recall across all untested pairs in the Restudy group ($M = 0.32$, $SD = 0.19$), $t(97) = 4.56$, $p < .001$, $d = 0.92$, $BF_{10} = 1182.75$. However, the comparison test for retrieval-induced forgetting in this and the previous experiment is not necessarily a fair comparison, as exposure to untested items is greater in the Restudy group, where every pair in the interim tasks is reviewed a second time. This is addressed in Experiment 3.

Experiment 3

In Experiment 2, we observed a benefit in new learning using partial distributed tests. However, the experiment was likely underpowered due to a limited sample size. Experiment 3 therefore aimed to replicate the critical effect with a larger sample size. In addition, the true extent of this effect may be underestimated, as the comparison group restudied every pair, and as such, exposure to study pairs differed between groups. Stated differently, performance in the Restudy group may be boosted in a way that creates an unfair comparison with the test groups. Experiment 3 aimed to provide a fairer control group, using a Distributed-Restudy group, where the structure of restudy blocks was identical to that in the Distributed-Test group. That is, participants in the Distributed-Restudy group did not restudy every pair, and some restudied pairs were from prior lists. The design is shown in Figure 5. This also provides an improved test of retrieval-induced forgetting, as we can compare cumulative test performance for untested pairs in the Distributed-Test group with un-restudied pairs in the Distributed-Restudy group, matching exposure.

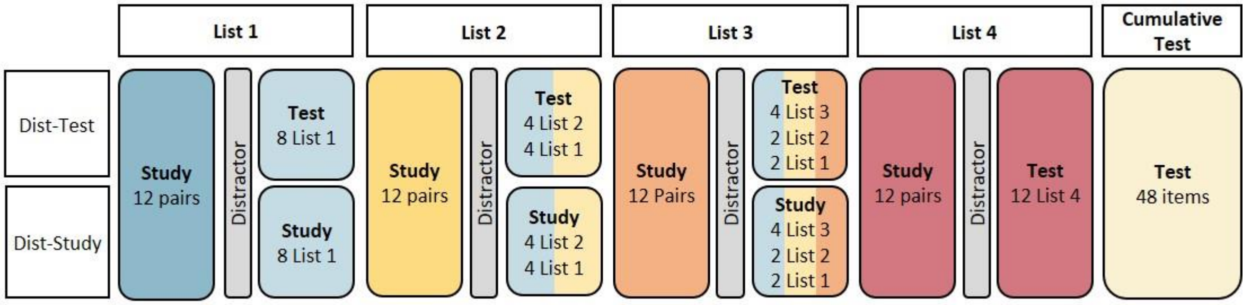

Figure 5. Schematic of the design of Experiment 3.

Method

Transparency and openness

The experiment's design, hypotheses and analysis plan were pre-registered at OSF, available at osf.io/c4vdh

Participants

Experiment 2 indicated an effect size of 0.36 in criterial test recall comparing the Distributed-Test and Restudy groups. A power analysis run in G*Power indicated a sample size of 133 participants per group was required to detect an effect of this size with 90% power. We recruited participants until we reached a total of 133 participants per group after exclusions. Data from a total of 276 participants were collected using Prolific. According to the pre-registered criteria, three participants were excluded for indicating they had taken notes during the experiment, seven for having prior knowledge of the Euskara language, and two participants for having a mean study time of less than 500 ms in List 4. Two participants fell into two of these categories, such that in total, data from 10 participants were excluded, giving a final sample of 266 participants (164 female, mean age = 36.15, *SD* = 13.76).

Materials

The materials were identical to those used in Experiment 2.

Procedure

The experiment used the same task procedure as Experiment 2 with some modifications. The Full-Test group was not included and the cumulative test occurred immediately after the criterial test within the same session. The design of the Distributed-Test group was identical to that in Experiment 2. The Distributed-Restudy group received the same distribution of pairs as the Distributed-Test group, however after each list, participants reviewed the pairs for up to 30 seconds each.

Results & Discussion

Criterial test

A one-tailed *t*-test indicated significantly greater recall in the Distributed Test group than the Distributed Restudy group, *t*(264) = 2.71, *p* = .004, *d* = 0.33, BF_{10} = 8.52 (see Figure 6A). Another one-tailed *t*-test showed that there were also significantly fewer intrusions in the Distributed Test than the Distributed Restudy group, *t*(264) = 5.07, *p* < .001, *d* = 0.62, BF_{10} = 35350.51 (see Figure 6B).

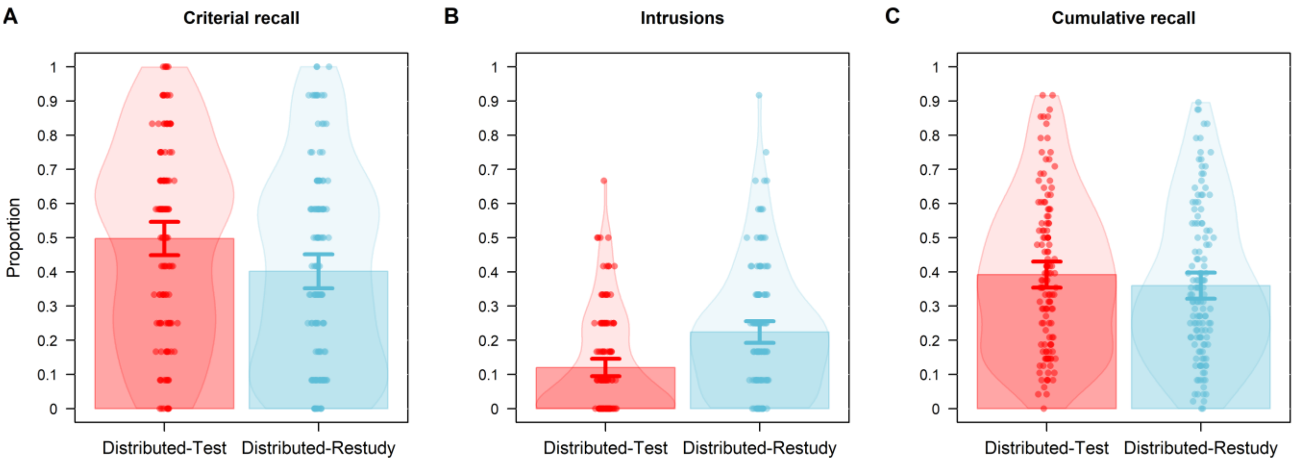

Figure 6. Results from Experiment 3. A: proportion of correct recall in the criterial test. B: proportion of intrusions in the criterial test. C: proportion of correct recall in the cumulative test. Error bars represent 95% confidence intervals.

Cumulative test

In this experiment, there was no benefit in an immediate cumulative test for the Distributed Test group overall. A one-tailed *t*-test showed that there was no significant difference in recall between the two groups, *t*(264) = 1.18, *p* = .120, *d* = 0.145, BF_{10} = 0.46 (see Figure6C).

A mixed effects logistic regression was conducted with Interim Task and number of times pairs were tested or restudied as fixed factors, and participant and pair as random factors to predict cumulative test recall for words from Lists 1-3[6]. The number of times a pair was included in interim tasks (either tested or restudied) was a significant predictor of correct recall in the cumulative test, *B* = 0.87, *SE* = 0.05, *z* = 16.57, *p* < .001, *OR* = 2.38, 95% CI [2.15,2.65]. Interim Task was also a significant predictor, *B* = 0.44, *SE* = 0.18, *z* = 2.43, *p* = .015, *OR* = 1.55 95% CI [1.08,2.22], and there was a significant interaction between Interim Task and number of times pairs were included, *B* = -0.47, *SE* = 0.07, z = -6.55, *p* < .001, *OR* = 0.62, 95% CI [0.54,0.71].

Figure 7A shows the proportion of correct recall in the cumulative test based on Interim Task and number of times pairs were included in interim tasks, and Figure 7B shows the regression function from the logistic regression analysis. Although it appears somewhat counterintuitive that the number of times pairs are included in interim tasks has greater influence on the Distributed-Restudy than the Distributed-Test group, part of the explanation is that test trials are much more beneficial if they include correct recall. Figure 4C shows the proportion of correct recall in the cumulative test, split by the number of times pairs were included in interim tasks, and data in the Distributed-Test group split according to whether the target was correctly recalled at least once in the interim tests, or recall was always incorrect. The figure illustrates that recalling a tested target correctly at least once in the interim tests improves cumulative test recall relative to restudying, and that pairs require multiple restudy opportunities to reach a similar level of cumulative test recall performance.

[6] In the analyses of cumulative test performance reported above for Experiments 1 and 2, we included data from all four lists as the focus was on the relationship between number of tests and cumulative test recall in the partial test groups. Experiment 3, in contrast, permits us to additionally make comparisons between the Distributed-Test and Distributed-Restudy group. For that reason, we exclude List 4 items from these analyses of cumulative test performance, as all items on this list received the same treatment (i.e., a test).

The left panel of Figure 7C shows cumulative test recall for items that were untested in the Distributed-Test group, and pairs that were not included in the interim restudy tasks in the Distributed-Restudy group (un-restudied items). There was no evidence of retrieval-induced forgetting, as a two-tailed *t*-test indicated there was greater recall of untested pairs in the Distributed-Test group (*M* = .32, *SD* = .25) compared to un-restudied pairs in the Distributed-Restudy group (*M* = .26, *SD* = .21), *t*(264) = 2.10, *p* = .037, *d* = 0.26, BF_{10} = 1.08. Thus, if anything, testing benefitted rather than harmed untested pairs (Rowland & DeLosh, 2014).

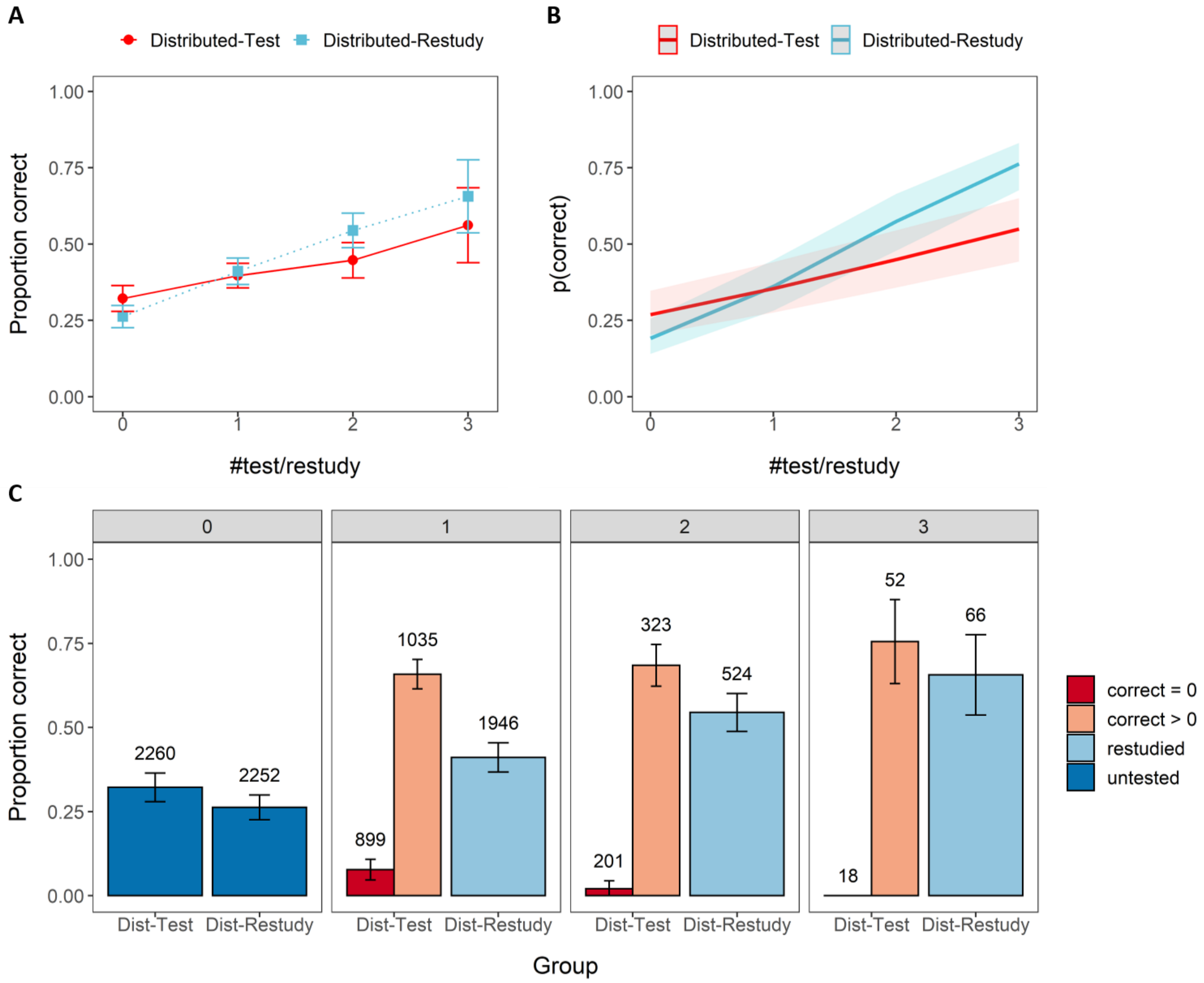

Figure 7. A: Proportion of correct recall of List 1-3 translations in the cumulative test by number of times items were included in interim tasks in the Distributed-Restudy and Distributed-Test groups. B: Logistic regression functions indicating the probability of correct recall of List 1-3

pairs by the number of times pairs were included in interim tasks. C: Proportion of correct recall in the cumulative test, where each panel indicates the number of times items were included in interim tasks, and data in the Distributed-Test group are split based on whether participants correctly recalled the target at least once, or never recalled the target correctly in the interim tests. Bar labels indicate the number of items in each condition collapsed across all participants. The left panel indicates the test of retrieval induced forgetting, comparing untested and un-restudied pairs.

Experiment 4

In Experiment 4, we aimed to further test the finding that partial testing does not reduce the forward testing effect, and determine whether testing even fewer pairs is still sufficient to enhance new learning in a high-powered experiment. To achieve this aim, we varied both interim task (test or restudy), and the number of pairs included in the interim task (12, 8, or 4 pairs). The design is illustrated in Figure 8.

Method

Transparency and openness

The experiment's design, hypotheses and analysis plan were pre-registered at OSF, available at osf.io/3pkj4.

Participants

We planned to analyse the data with a between-subjects ANOVA, and pairwise comparisons of Interim Task within each Number of Pairs condition, and ran power analyses for both in G*Power, based on the mean effect size from the previous two experiments of 0.35. The ANOVA required a total sample size of 346 participants to detect main effects and an interaction based on a converted effect size $f = 0.175$. The pairwise independent samples t-tests indicated a sample size of 141 participants per group was required to detect a forward testing effect with

90% power, giving a total of 846 participants. This effect size is based on the mean effect size in the previous two experiments.

Data from a total of 868 participants were collected, sequentially allocated to each of the six groups based on the order they began the experiment. Eleven participants indicated they had taken notes during the experiment, seven that they had prior knowledge of the Euskara language, and four had a mean study time of less than 500 ms in List 4, and were excluded from the analyses. In total, data from 22 participants was excluded, giving a final sample of 846 participants (164 female, mean age = 36.15, *SD* = 13.76). There were 139 participants in the Test-12 pair group, 139 in the Restudy-12 pair group, 140 in the Test-8 pair group, 143 in the Restudy-8 pair group, 138 in the Test-4 pair group, and 147 in the Restudy 4-pair group.

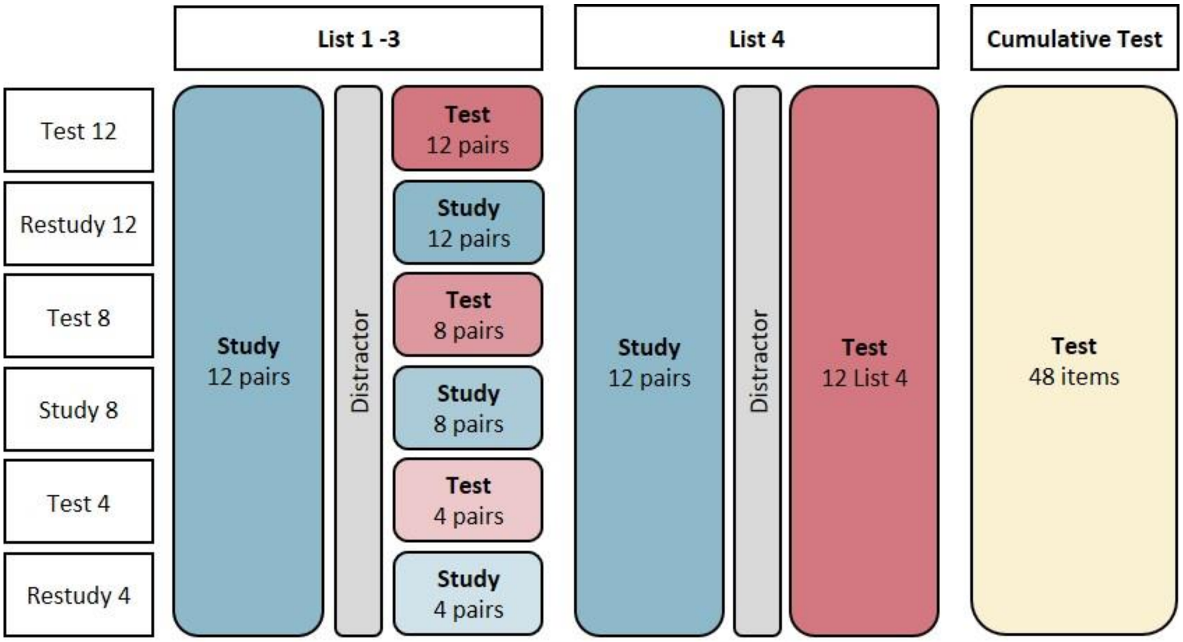

Figure 8. Schematic of the design in Experiment 4.

Design

The experiment used a 2 (Interim Task: Test vs Restudy) x 3 (Number of Pairs: 12 vs 8 vs 4) between-subjects design, which is detailed in Figure 8. Participants in the Test conditions received a test after studying each list, while the Restudy groups reviewed the pairs a second time. Participants were either tested on or reviewed 12, 8, or 4 pairs from the previous list, depending on the Number of Pairs condition.

Procedure

The experiment used the same task as in Experiment 2 and 3. After studying words from each of the first three lists, participants completed a distractor maths task for 1 minute before receiving a test on pairs from the previous list, or reviewing pairs from that list a second time. Participants in the 12-pair conditions saw all pairs from the previous list, participants in the 8-pair condition saw 8 randomly selected pairs from the previous list, and participants from the 4-pair condition saw 4-pairs from the previous list. Following study and the distractor task for List 4, all six groups were tested on all 12 List 4 pairs, as well as a cumulative test on all 48 pairs in random order.

Results & Discussion

Criterial test

Proportion of correct recall in the criterial test is shown in the left column of Figure 9. We first ran a 2 (Interim Task: Test vs Restudy) x 3 (Number of Pairs: 4 vs 8 vs 12) ANOVA on criterial test scores, which showed a significant main effect of Interim Task, indicating a forward testing effect, $F(1,840) = 20.35$, $p < .001$, $\eta^2_p = .024$, $BF_{10} = 1563$. There was no significant main effect of Number of Pairs, $F(2,840) = 2.18$, $p = .114$, $\eta^2_p = .005$, $BF_{01} = 8.78$, and crucially no significant interaction between Interim Task and Number of Pairs, $F(2,840) = 0.93$, $p = .396$, η^2_p

= .002, BF_{excl} = 15.42. Planned interaction contrasts showed no difference in the strength of the forward testing effect (the difference in List 4 recall between the Test and Restudy groups) for the full (12-pair) conditions compared to the partial conditions combined (4 and 8-pairs), $t(840) = 1.35$, $p = .177$, $BF_{excl} = 3.78$[7], and no difference in the magnitude of the forward testing effect between the two partial groups, $t(840) = 0.17$, $p = .862$, $BF_{excl} = 7.45$[8]. To determine whether there was a significant effect in each Number of Pairs condition, we ran three separate one-tailed *t*-tests. There was a significant FTE in the 12-pair, $t(276) = 3.60$, $p < .001$, $d = 0.43$, BF_{10} = 114.83 , 8-pair, $t(281) = 2.18$, $p = .015$, $d = .259$, $BF_{10} = 2.43$, and 4-pair, $t(283) = 1.99$, p = .024, $d = .235$, $BF_{10} = 1.64$ conditions.

Proportion of intrusions in the criterial test is shown in the middle column of Figure 9. There was a significant main effect of Interim Task on the number of intrusions, $F(1,840)$ = 114.57, $p < .001$, $\eta^2_p = .120$, $BF_{10} = 1.04 \times 10^{22}$, but there was no main effect of Number of Pairs, $F(2,840) = 0.73$, $p = .484$, $\eta^2_p = .002$, $BF_{01} = 35.14$, nor an interaction between these factors, $F(2,840) = 0.56$, $p = .573$, $\eta^2_p = .001$, $BF_{excl} = 22.72$. Planned interaction contrasts showed no difference in the effect of Interim Task between full and partial conditions combined, $t(840)$ = .229, $p = .819$, $BF_{01} = 8.46$, or the two partial conditions, $t(840) = 1.03$, $p = .303$, $BF_{01} = 4.67$.[9]

[7] Bayes factors for contrasts were calculated using the method proposed by Morey (2015).

[8] An exploratory analysis of criterial test recall also showed no significant difference in the magnitude of the forward testing effect between the 12- and 8-pairs conditions, $F(1,557) = 1.15$, $p = .285$, $\eta^2_p = .002$, $BF_{excl} = 4.42$, or the 12- and 4-pairs conditions, $F(1,559) = 1.58$, $p = .209$, $\eta^2_p = .003$, $BF_{excl} = 3.86$. Thus, although the effect sizes are somewhat smaller in the partial conditions, there is no significant difference in the magnitude of the forward testing effect between either partial condition and the 12-pair condition.

[9] An exploratory analysis of prior-list intrusions also showed no significant interaction between Interim Task and Number of Pairs when comparing the 12- and 8-pairs conditions, $F(557)$ = 0.10, $p = .748$, $\eta^2_p < .001$, $BF_{excl} = 7.00$, or the 12- and 4-pairs conditions, $F(1,559) = 0.48$, p = .489, $\eta^2_p < .001$, $BF_{excl} = 5.82$.

One-tailed pairwise comparisons revealed a significant difference in prior list intrusions between Test and Restudy groups in the 12-pair, $t(276) = 7.25$, $p < .001$, $d = .748$, $BF_{10} = 1.08 \times 10^7$, 8-pair $t(281) = 7.25$, $p < .001$, $d = .862$, $BF_{10} = 3.49 \times 10^9$ and 4-pair, $t(283) = 5.17$, $p < .001$, $d = .613$, $BF_{10} = 57659$, conditions.

Cumulative test

Proportion of correct recall in the cumulative test is shown in the right column of Figure 9. There was no significant effect of Interim Task on cumulative test recall, $F(1,840) = .20$, $p = .657$, $\eta^2_p < .001$, $BF_{01} = 11.55$, and no interaction between Interim task and Number of Pairs, $F(2,840) = 0.47$, $p = .626$, $\eta^2_p = .001$, $BF_{excl} = 24.98$. However, there was a significant main effect of Number of Pairs, $F(2,840) = 11.80$, $p < .001$, $\eta^2_p = .027$, $BF_{10} = 1079.60$. Recall was significantly higher in the full conditions than the partial conditions combined, $t(840) = 3.81$, $p < .001$, $BF_{10} = 91.25$ and significantly higher in the 8-pair than 4-pair conditions, $t(840) = 3.81$, $p < .001$, $BF_{10} = 10.16$.

One-tailed *t*-tests showed no significant difference between the test and restudy groups in the 12-pair, $t(276) = 0.98$, $p = .164$, $d = .118$, $BF_{01} = 2.89$, 8-pair, $t(281) = 0.19$, $p = .574$, $d = .022$, $BF_{01} = 8.79$, and 4-pair, $t(283) = 0.10$, $p = .538$, $d = .011$, $BF_{01} = 8.25$, conditions.

To test retrieval-induced forgetting, we compared recall for untested pairs in the partial test conditions versus un-restudied pairs in the partial restudy conditions. A 2 (Interim Task: Test vs Restudy) x 2 (Number of Pairs: 4 vs 8) ANOVA showed no significant effect of Interim Task, $F(1,564) = 0.72$, $p = .397$, $\eta^2_p = .001$, $BF_{01} = 7.63$, no significant effect of Number of Pairs, $F(1,564) = 1.98$, $p = .160$, $\eta^2_p = .003$, $BF_{01} = 4.09$, and no significant interaction, $F(1,564) = .09$, $p = .761$, $\eta^2_p < .001$, $BF_{excl} = 7.34$, demonstrating no evidence of retrieval-induced forgetting.

Two-tailed *t*-tests found no significant effect of Interim Task in the 8-pair condition, *t*(281) = 0.74, *p* = .461, *d* = .088, BF_{01} = 5.90, or 4-pair condition, *t*(283) = 0.44, *p* = .664, *d* = .052, BF_{01} = 7.01 (this analysis cannot be applied to the 12-pair condition).

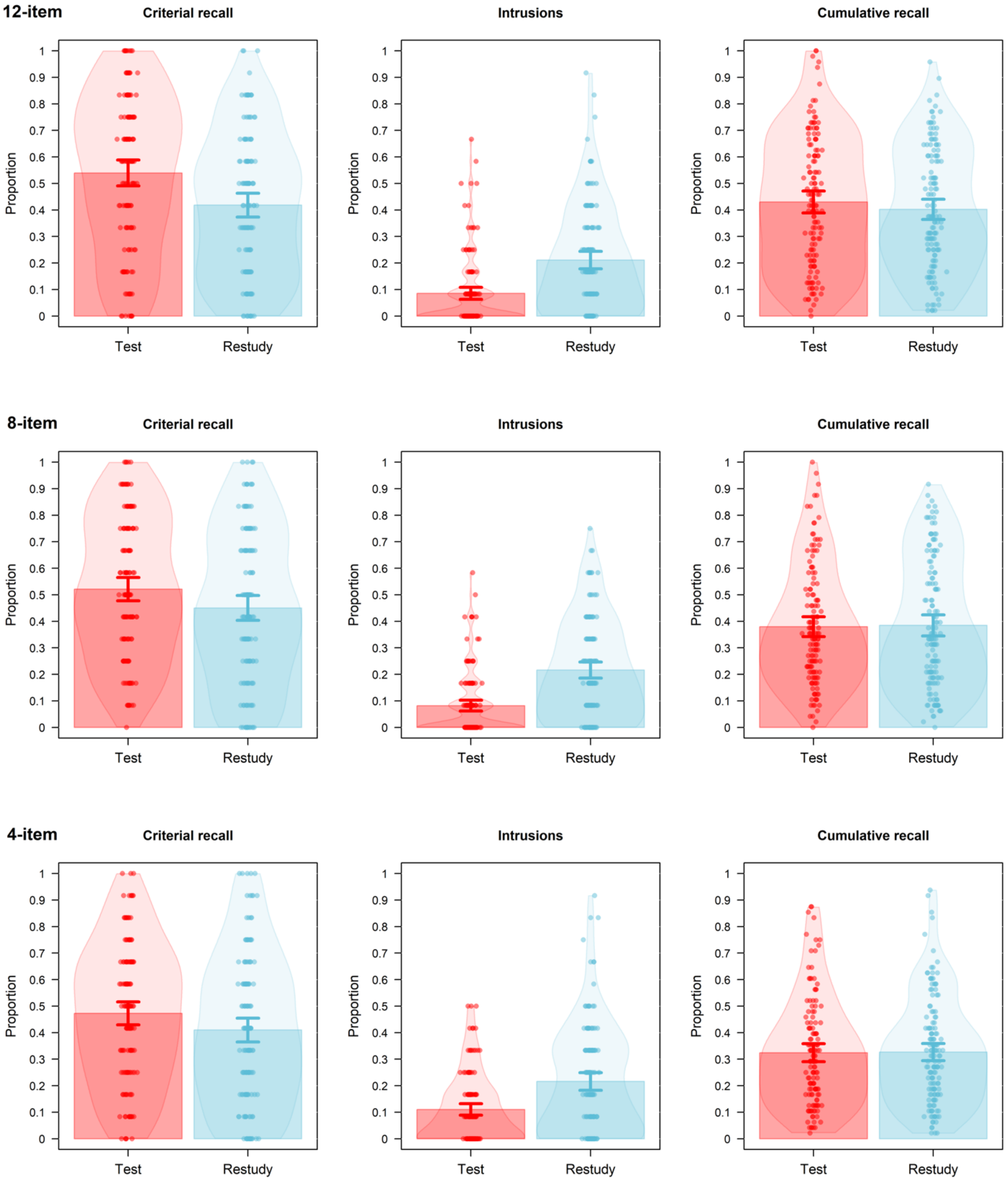

Figure 9. Results from Experiment 4 in the 12-pair (top panel), 8-pair (middle panel) and 4-pair (bottom panel) conditions. Left column shows proportion of correct recall in the criterial test, middle column shows proportion of intrusions in the criterial test, and right column shows proportion of correct recall in the cumulative test. Error bars represent 95% confidence intervals.

Prior-list intrusions and release from PI

Intrusions have been used in previous research to indicate the degree of proactive interference (e.g., Aslan & Bäuml, 2016; Bufe & Aslan, 2018; Nunes & Weinstein, 2012; Pastötter & Bäuml, 2014; Pastötter et al., 2013; Pierce et al., 2017; Szpunar et al., 2008l Weinstein et al., 2011; Weinstein et al., 2014; Yang et al., 2017; Yang et al., 2019). Yang et al. (2022) found that prior list intrusions mediated the effect of interim task on criterial test recall with single-item Chinese word lists, indicating a contribution of release from PI to the forward testing effect. Here, we ran conceptually the same analyses for all four experiments on each pairwise group comparison. Mediation analyses were run using the R *mediation* package (Imai, Keele, & Yamamoto, 2010), with List 4 recall as the dependent variable, Interim Task (Full-Test vs Restudy; Distributed-Test vs Restudy) as the independent variable, and prior list intrusions in List 4 recall as the mediator, with 1000 bootstrap samples. In all cases, intrusions were a significant mediator of the effect of Interim Task on recall, with prior list intrusions accounting for 54.0-99.8% of the effect of testing on recall. Moreover, none of the direct effects were statistically significant, suggesting near-complete mediation by intrusions. The full results from the mediation analyses are shown in Table 1.

Table 1.

Mediation analyses

		Full-Test vs. Restudy			Partial-Test vs. Restudy					
Exp	Mediation model	*β*	*95% CI*	*p*	*β*	*95% CI*	*p*			
1	Total effect	0.134	[0.01, 0.26]	.040*	0.221	[0.10, 0.33]	<.001*			
	Direct effect	0.002	[-0.10, 0.12]	.970	0.102	[-0.004, 0.20]	.064			
	Indirect effect	0.132	[0.06, 0.21]	<.001*	0.119	[0.06, 0.19]	<.001*			
	Proportion explained	0.988	[0.39, 3.59]	0.04 *	0.540	[0.27, 1.03]	<.001*			
2	Total effect	0.155	[0.05, 0.27]	.004*	0.112	[0.002, 0.23]	.046*			
	Direct effect	0.058	[-0.05, 0.17]	.258	0.035	[-0.07, 0.13]	.524			
	Indirect effect	0.097	[0.05, 0.15]	<.001*	0.077	[0.03, 0.13]	.002*			
	Proportion explained	0.628	[0.31, 1.82]	.004*	0.685	[0.10, 4.03]	.048*			
3	Total effect				0.096	[0.02, 0.16]	.008*			
	Direct effect				0.013	[-0.05, 0.07]	.706			
	Indirect effect				0.083	[0.05, 0.12]	<.001*			
	Proportion explained				0.867	[0.49, 2.75]	.008*			
		12-pair			8-pair			4-pair		
4	Total effect	0.121	[0.06,0.19]	<.001*	0.071	[0.01, 0.13]	0.026*	0.063	[0.003, 0.12]	0.038 *
	Direct effect	0.016	[-0.04,0.07]	.570	-0.040	[-0.10, 0.03]	0.218	-0.021	[-0.08, 0.03]	0.458
	Indirect effect	0.105	[0.08,0.14]	<.001*	0.111	[0.08, 0.15]	<.001*	0.086	[0.06, 0.12]	<.001*
	Proportion explained	0.864	[0.57,1.66]	<.001*	1.561	[0.70, 8.72]	0.026	1.359	[0.61, 6.14]	0.038 *

Note: Asterisks indicate a significant effect, *p* < .05. The indirect effect represents the portion of the relationship between interim task and recall that is mediated by intrusions. The direct effect reflects the portion of the relationship between interim task and recall that is not mediated by intrusions. The total effect indicates the combined direct and indirect effects.

General Discussion

The forward testing effect has clear applications in educational practice, where maintaining study effort, learning efficiency, and retention of learned materials are important (Szpunar et al., 2013). In this study, we aimed to determine whether partial and distributed tests – probably more representative of classroom practice than full tests – are sufficient to produce the forward testing effect, and whether they are detrimental to untested material.

We consistently found benefits in new learning with partial tests (Experiment 1 & 4) and partial distributed tests (Experiments 2 & 3) using multi-list, paired associate designs with both face-name and foreign-language vocabulary learning. All four experiments replicated the FTE, with enhanced learning in a full test group relative to a restudy group. Experiment 1 showed that a partial test, which only included half the pairs from the previous list, was effective in enhancing criterial test learning. Experiment 2 found that partial distributed tests, which included some pairs from the just-studied list, as well as some pairs from prior lists, also enhanced learning relative to a full restudy group. Experiment 3 showed that this condition was also effective compared to a distributed restudy group. Experiment 4 demonstrated that testing 8 or 4 (out of 12) pairs could generate a forward testing effect when compared to restudying 8 or 4 pairs, respectively. Although there was some decrease in the effect size of the forward testing effect with fewer tested items in Experiment 4, in all cases, the magnitude of the effects produced by partial and/or distributed tests did not differ significantly from the effect produced by a full test. We did not directly compare or separate the effects of partial versus distributed testing, and so it is not clear whether there are differences in the strength of forward testing effects based on test pair distribution, which should be an avenue of future research.

Nevertheless, from the available data, there is no reason to assume that partial, or partial and distributed testing is inferior to full testing.

Prior list intrusions were greater in restudy conditions than either full or partial test groups in all four experiments. An important theoretical contribution is the finding that PI, as measured by prior list intrusions, substantially mediated the effect of testing on recall, providing support for release-from-PI theories (Szpunar et al., 2013). Although results were somewhat mixed, on the whole, we also observed lower response latency in test and partial test conditions relative to restudy, which also suggests less PI in these groups (see Supplemental Material). On the one hand, one might assume that partial and distributed tests would be less effective in reducing PI, as they may provide less differentiation between old and new lists. On the other, it is possible that any amount of retrieval from long-term memory, even if not from the just-studied list, is sufficient to reduce PI (see Pastötter et al., 2011). The results from the partial test conditions appear to indicate the latter. Intrusions mediated the effect of interim task on criterial test recall for both full- and partial-test groups, a novel result which conceptually replicates findings from Yang et al. (2022) and extends them to paired-associate learning and partial test conditions.

While the current experiments were not designed to test different theories of the forward testing effect, the results might also indicate that partial tests are sufficient to induce a context change that allows reset of encoding, which maintains the effectiveness of encoding (Pastötter et al., 2011), or to allow strategy changes that facilitate encoding or retrieval (Cho et al., 2017; Soderstrom & Bjork, 2014), although the experiments do not allow us to measure the effect of partial tests on these processes directly. Partial tests also increase test expectancy relative to restudy, to a similar degree as full tests (see Supplemental Material for Experiment 1). Increased test expectancy could lead to greater motivation or attention when encoding new material (Weinstein et al., 2014).

It is theoretically significant that partial tests induced a forward testing effect which was not offset by any detectable costs in the cumulative tests, given the superficial resemblance between the procedure used in these experiments and those employed in research on retrieval-induced forgetting (Bäuml & Kliegl, 2017). Typically, the literature on retrieval-induced forgetting compares non-tested pairs within a tested category with non-tested pairs from a non-tested category to determine whether retrieval-induced forgetting or facilitation occurs. The materials used in the current tasks were not separated into categories. We therefore assessed retrieval induced forgetting between subjects, comparing non-tested pairs in the test group with non-tested pairs in the study group. In Experiment 3 and 4, we compared non-tested and non-restudied pairs, to better match exposure.

Chan (2009) demonstrated that low-integrative encoding and a short recall delay resulted in retrieval-induced forgetting, while high-integrative encoding and a 24-hour recall delay resulted in retrieval-induced facilitation for both prose material and simple propositional sentences. Integration of material is suggested to eliminate retrieval-induced forgetting by reducing retrieval competition between studied items (Anderson, 2003), and a delay between study and recall reduces retrieval-induced forgetting as this response competition decreases over time (MacLeod & Macrae, 2001). The paired-associate materials used in the current study are naturally low-integrative, and Experiments 1 and 3 used a short delay before final recall. Thus, partial testing did not impair recall of untested targets under conditions in which retrieval-induced forgetting would, if anything, be most expected. The only evidence for retrieval-induced forgetting was in Experiment 2, but participants in the Restudy group had greater exposure to untested pairs than in the Distributed-Test group. Notably, this was the only experiment with a 24-hour delay prior to the cumulative test. In Experiment 3, where exposure to pairs was matched, there was no evidence for a deficit in retrieval for untested pairs compared to the Restudy group, using a large sample size. Instead, there was some evidence for facilitation of recall of untested pairs (though the Bayesian evidence was weak). Thus, it is clear that the forward testing effect can occur independently of retrieval-induced forgetting, at least for the paired-associate materials used here. Future research should examine whether these results can be replicated with more complex study materials, and materials with greater interference between retrieved and non-retrieved pairs.

Effects of testing on overall cumulative test recall were mixed. In most cases, there was no benefit of testing in the cumulative test, with the exception of the Full-Test group in Experiment 2. This is consistent with some equivalent null effects in cumulative tests reported elsewhere (e.g., Wissman & Rawson, 2015). Interpretation of the cumulative test data as a whole is quite complex, involving influences of both proactive and retroactive interference, as well as both forward- and backward-testing effects. With the exception of Experiment 2, which used a 24-hr delay, the experiments had a very brief lag before the cumulative test. This may have contributed to the null effects, as testing effects tend to be stronger with longer delays (Roediger & Karpicke, 2006).

The current results are useful for translation of the FTE into classroom settings, as they demonstrate that testing of all learned material is not necessary to produce forward beneficial effects of testing, and that including pairs from prior lists does not interfere with this beneficial effect, or increase intrusions from prior lists. In addition, Experiments 3 and 4 clearly demonstrate that the benefits of partial testing are not offset by recall deficits for untested pairs.

Acknowledgements

This work was supported by the Economic and Social Research Council [grant number ES/S014616/1] and the Natural Science Foundation of China [grant number 32000742]. Preregistrations and data, not including analysis code are available at the Open Science Framework, https://osf.io/3pkj4/?view_only=678a6b8440f74194b9a206c3265660bd. Study materials are available in the Supplemental Material.

References

Aslan, A., & Bäuml, K. H. T. (2016). Testing enhances subsequent learning in older but not in younger elementary school children. *Developmental Science, 19*, 992-998. https://doi.org/10.1111/desc.12340

Bauml, 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. https://doi.org/ 10.1016/j.jml.2012.07.006

Bäuml, K. -H. T., & Kliegl, O. (2017). Retrieval-induced remembering and forgetting. In J. T. Wixted, & J. H. Byrne (Eds.) *Cognitive Psychology of Memory, Vol. 2 of Learning and Memory: A Comprehensive Reference, 2nd edn* (pp. 27–51). Oxford: Academic Press.

Bolte, C. M. (2019). *When is test-potentiated learning item-specific versus generalized?* (Masters Dissertation). State University of New York at Albany, ProQuest Dissertations Publishing, 13813187.

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

Chan, J. C. (2009). When does retrieval induce forgetting and when does it induce facilitation? Implications for retrieval inhibition, testing effect, and text processing. *Journal of Memory and Language*, *61*(2), 153-170. https://doi.org/10.1016/j.jml.2009.04.004

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

Cho, K. W., Neely, J. H., Crocco, S. & Vitrano, D. (2017). Testing enhances both encoding and retrieval for both tested and untested items. *Quarterly Journal of Experimental Psychology, 70*, 1211–1235. https://doi.org/10.1080%2F17470218.2016.1175485

DeBruine, L., & Jones, B. (2017). *Face Research Lab London Set (Version 3)*. https://doi.org/10.6084/m9.figshare.5047666.v3.

Don, H., & Shanks, D. (2022, February 8). Do partial tests enhance new learning? Retrieved from osf.io/3pkj4

Faul, F., Erdfelder, E., Lang, A. G., & Buchner, A. (2007). G* Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*(2), 175-191. https://doi.org/10.3758/BF03193146

Imai, K., Keele, L., Tingley, D., & Yamamoto, T. (2010). Causal mediation analysis using R. In *Advances in social science research using R* (pp. 129-154). Springer, New York, NY.

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. https://doi.org/10.1037/xap0000087

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

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, and Cognition, 40,* 1787-1794.

MacLeod, M. D., & Macrae, C. N. (2001). Gone but not forgotten: The transient nature of retrieval-induced forgetting. *Psychological Science, 12*, 148–152. https://doi.org/10.1111%2F1467-9280.00325

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. 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*, 286. https://dx.doi.org/10.3389%2Ffpsyg.2014.00286

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

Pastötter, B., Weber, J., & Bäuml, K.-H. T. (2013). Using testing to improve learning after severe traumatic brain injury. *Neuropsychology, 27*(2), 280–285. http://dx.doi.org/10.1037/a0031797

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

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. https://doi.org/10.1111%2Fj.1745-6916.2006.00012.x

Rouder, J. N., Morey, R. D., Verhagen, J., Swagman, A. R., & Wagenmakers, E.-J. (2017). Bayesian analysis of factorial designs. *Psychological Methods, 22*(2), 304–321. https://doi.org/10.1037/met0000057

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

Rowland, C. A., & DeLosh, E. L. (2014). Benefits of testing for nontested information: Retrieval-induced facilitation of episodically bound material. *Psychonomic Bulletin & Review, 21*, 1516-1523. https://doi.org/10.3758/s13423-014-0625-2

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

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

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. https://doi.org/10.1073/pnas.1221764110

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

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

Wagenmakers, E.J. et al. (2018). Bayesian inference for psychology Part II: Example applications with JASP. *Psychonomic Bulletin & Review, 25*, 58-76. doi: 10.3758/s13423-017-1323-7

Wang, T., Yang, C., & Zhong, N. (2020). Forward testing effect on new learning in older adults. *Acta Psychologica Sinica, 52*, 1266–1277. https://doi.org/10.3724/SP.J.1041.2020.01266

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

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

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

Wissman, K. T., & Rawson, K. A. (2015). Grain size of recall practice for lengthy text material: Fragile and mysterious effects on memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 41*, 439–455.

Wixted, J. T., & Rohrer, D. (1993). Proactive interference and the dynamics of free recall. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 19*, 1024. https://doi.org/10.1037/0278-7393.19.5.1024

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. https://psycnet.apa.org/doi/10.1037/edu0000320

Yang, C., Luo, L., Vadillo, M. A., Yu, R., & Shanks, D. R. (2021). Testing (quizzing) boosts classroom learning: A systematic and meta-analytic review. *Psychological Bulletin*. Advance online publication. https://doi.org/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. https://doi.org/10.1037/xap0000122

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

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. https://psycnet.apa.org/doi/10.1037/edu0000031