# The Forward Testing Effect on Self-Regulated Study Time Allocation and Metamemory Monitoring

Chunliang Yang, Rosalind Potts, and David R. Shanks

University College London

## Author Note

This research was supported by the China Scholarship Council (CSC).

Correspondence concerning this article should be addressed to David R. Shanks, Division of Psychology and Language Sciences, University College London, 26 Bedford Way, London WC1H 0AP. Email: d.shanks@ucl.ac.uk.

All experimental data have been made publicly available via the Open Science Framework (OSF) at https://osf.io/rqm6g/.

**Abstract**

The forward testing effect describes the finding that testing of previously studied information potentiates learning and retention of new information. Here we asked whether interim testing boosts self-regulated study time allocation when learning new information and explored its effect on metamemory monitoring. Participants had unlimited time to study five lists of Euskara-English word pairs (Experiment 1) or four lists of face-name pairs (Experiment 2). In a No Interim Test group which was only tested on the final list, study time decreased across successive lists. In contrast, in an Interim Test group, which completed a recall test after each list, no such decrease was observed. Experiments 3 and 4 were designed to investigate the forward testing effect on metamemory monitoring and found that this effect is associated with metacognitive insight. Overall, the current study reveals that interim tests prevent the reduction of study time across lists and that people's metamemory monitoring is sensitive to the forward benefit of interim testing. Moreover across all four experiments, the Interim Test group was less affected by proactive interference in the final list interim test than the No Interim Test group. The results suggest that variations in both encoding and retrieval processes contribute to the forward benefit of interim testing.

***Keywords*:** forward testing effect; self-regulated learning; encoding; retrieval; metamemory monitoring

With the increasing popularity and availability of free online courses and learning aids, self-regulated learning is taking place more and more outside of the formal classroom (Bjork, Dunlosky, & Kornell, 2013). To use these opportunities effectively, learners must understand how to regulate their behaviour to optimize learning, comprehension, and knowledge transfer. However, recent studies reveal that we are far from being sophisticated learners (for a review, see Bjork et al., 2013). Therefore, self-regulated learning has become a significant focus of theoretical and empirical research for both psychologists and educators.

A few studies have been conducted employing interim tests to optimize self-regulated learning of previously studied or tested information (Karpicke, 2009; Soderstrom & Bjork, 2014). But no research has yet been undertaken employing interim tests to optimize self-regulated learning of new information. One aim of the current study is to fill this gap. Specifically, we explored how interim tests influence subsequent self-regulated study time allocation when learning new information.

**Backward testing effect**

In educational settings, testing is usually regarded as an evaluative instrument to assess learning and comprehension. A large body of research has supplied convincing evidence that testing is also an effective instrument to facilitate long term retention (for a review, see Roediger & Karpicke, 2006a). The common finding that retrieval of previously studied information enhances its retention by comparison with restudying that information or doing nothing was first explored over 100 years ago (Abbott, 1909) and is usually termed the *testing effect* (for review, see Roediger & Karpicke, 2006a; Roediger, Putnam, & Smith, 2011). We use the term *backward testing effect* for this phenomenon, following Pastötter and Bäuml (2014). Researchers have suggested that retrieval practice (i.e., testing) engages deeper and more elaborative processing, which improves retrieval accessibility in a later test (Carpenter, 2009; Roediger & Karpicke, 2006a), a direct mechanism by which testing can enhance retention of the tested information (Roediger et al., 2011).

Testing can also enhance retention of tested information in some other, indirect, ways. For example, learners may take test results as feedback to diagnose the gap between their on-going learning status and their desired status and then regulate their subsequent learning to narrow this gap (Pyc & Rawson, 2010; Pyc & Rawson, 2012). Another indirect testing effect is that interim tests can improve subsequent encoding efficiency when the same material is restudied, a phenomenon termed the *potentiating effect of testing* (Arnold & McDermott, 2013; Izawa, 1969). For example, Pyc and Rawson (2012) had participants study Swahili-English word pairs. Participants were instructed to employ a keyword encoding strategy, generating and reporting a keyword to associate a Swahili word and its corresponding translation. In a test-restudy group, higher proportions of keyword shifts took place than in a restudy group, and higher proportions of keywords were modified following retrieval failure versus retrieval success. During retrieval attempts, participants evaluated the efficiency of their self-generated mediators and modified less effective keywords. Hence interim testing can facilitate subsequent re-encoding and render tested material more retrievable in future. Karpicke, Lehman, and Aue (2014) proposed that retrieval practice updates a given item's context so that that item is associated with multiple encoding and retrieval context cues, which facilitate its subsequent recall (for a review of the direct and indirect ways by which testing enhances retention, see Roediger et al., 2011).

**Forward testing effect**

Recent research has supplied evidence that testing of previously studied information from a given domain (i.e., a specific type of information) can also improve encoding and retention of new information from the same domain. Several terms have been used to refer to this effect[1]. In this study, we term it *the forward testing effect* in contrast to the better-known backward testing effect.

In Szpunar, McDermott, and Roediger's (2008) Experiment 1A, participants were instructed to study 5 18-word lists in anticipation of a cumulative test. In an Interim Test group, participants undertook a free recall test after studying each individual list. In a No Interim Test group, participants were only tested on List 5. In the List 5 interim test, the Interim Test group recalled more List 5 words and suffered less *proactive interference* (PI; i.e., mistakenly recalling words from prior lists) than the

No Interim Test group. In Experiment 2, five groups of participants were recruited. One group was tested on every list. The other 4 groups were only tested on one of Lists 2-5. The results showed that recall on a given list was always better when previous lists had been tested than they had not been tested. Moreover, with an increasing number of untested lists, interim test recall decreased and the amount of proactive interference increased. Thus the greater the number of previously untested lists prior to a tested list, the worse was interim recall on that test. This forward testing effect has been replicated with a range of materials, including words (Bäuml & Kliegl, 2013; Pastötter, Schicker, Niedernhuber, & Bäuml, 2011; Pastötter, Weber, & Bäuml, 2013), face-name pairs (Weinstein, McDermott, & Szpunar, 2011), online courses (Jing, Szpunar, & Schacter, 2016; Schacter & Szpunar, 2015; Szpunar, Khan, & Schacter, 2013), pictures (Pastötter et al., 2013), texts (Wissman, Rawson, & Pyc, 2011), and Swahili-English word pairs (Cho, Neely, Crocco, & Vitrano, 2016). This effect is not limited to heathy individuals, but also extends to people who suffer from severe traumatic brain injury (Pastötter et al., 2013).

Szpunar et al. (2008) proposed a retrieval theory to account for the forward testing effect, which postulates that it is caused by release from PI. Retrieval practice induces more substantial between-list context changes, and these in turn facilitate list discrimination at the time of recall and reduce interference. Put differently, the Interim Test group takes advantage of list-specific contexts at retrieval to limit the memory search set, which reduces PI and assists current list recall. Bäuml and Kliegl (2013) provided further evidence to support this contextual list segregation conjecture. They asked participants to study three lists of words. In the List 3 interim test, participants in the Interim Test group recalled more words than participants in the No Interim Test group, and the Interim Test group's response latencies were shorter than those in the No Interim Test group. Shorter response latencies imply a smaller memory search set, consistent with more effective discrimination between the target and non-target lists.

An alternative encoding theory postulates that interim testing makes subsequent encoding of new information as effective as the encoding of prior lists, while in the absence of interim tests, the encoding of new information deteriorates across lists. Pastötter et al. (2011) recorded participants'

brain activity while studying five 18-word lists. Electroencephalogram (EEG) results showed that alpha power, which is linked to reduced attention (Palva & Palva, 2007), increased across lists in the No Interim Test group but not in the Interim Test group. Also supporting the encoding account is evidence from Szpunar et al. (2013). They had participants study 4 segments of an introductory statistics video and measured participants' mind-wandering during encoding by asking them to report whether or not their attention was on-task. Participants in the No Interim Test group reported more mind-wandering than those in the Interim Test group. Similarly, Jing et al. (2016) found that participants in their Interim Test group reported fewer task-unrelated thoughts (zoning out) but more task-related thoughts (e.g., thoughts relating the course to their own life) than in the No Interim Test group while studying an online course. More task-unrelated thoughts lead to worse learning while more task-related thoughts are linked to better learning. Pastötter et al. (2011) proposed a specific mechanism to explain why interim testing prevents encoding deterioration across lists, namely that retrieval practice (interim testing) produces an internal context change which induces a reset of encoding and makes subsequent encoding of new information as effective as encoding of previous information. Weinstein, Gilmore, Szpunar, and McDermott (2014) proposed an alternative mechanism. They suggested that participants' expectancy of an immediate interim test in the Interim Test group remained constant or increased consistently across lists but decreased across lists in the No Interim Test group. Expecting an upcoming test forced participants to focus their attention and learning effort towards encoding new information.

These two possible mechanisms (the encoding and retrieval mechanisms) are not mutually exclusive and both may contribute to the forward testing effect. In the current study, by directly measuring study time allocation, we explore the contribution of variations in encoding processes to the forward testing effect. Moreover, by measuring the difference in PI between the Interim Test and No Interim Test groups, we explore the contribution of variations in retrieval processes to this effect.

**Self-regulated learning**

In some situations, learners can manage their learning in near-optimal ways to induce memory formation. For instance, Kornell and Metcalfe (2006) asked participants to choose which half

of a set of word pairs they preferred to restudy later. In the honouring condition, participants reviewed the pairs which were selected to be restudied. In contrast, participants in the dishonouring condition reviewed the pairs which they had not selected for restudy. In a later test, participants in the honouring condition significantly outperformed those in the dishonouring condition. This study revealed that people can manage their learning in a relatively effective way when their assessment of learning is accurate. Nonetheless, self-regulated learning does not always lead to better learning. In some situations, self-regulated learning impairs retention compared with experimenter-paced learning. For instance, Kornell and Bjork (2008) allowed some participants to remove some Swahili-English pairs from further study which they thought were well-studied and did not need further study, while others were not allowed to remove any pairs. Removing pairs from further study impaired retention, and Kornell and Bjork (2008) concluded that people tend to end learning prematurely before they reach the proximal learning region (Metcalfe & Kornell, 2005).

Recent research has employed interim tests to enhance self-regulated learning of previously studied or tested information (Soderstrom & Bjork, 2014). In Soderstrom and Bjork's (2014) Experiment 1, participants were asked to study a mixture of unrelated, forward-, and backward-related word pairs. For the unrelated pairs (e.g., *paper-ball*), there was no semantic association from the cue to target words and no association from the target to the cue words. For the forward-related pairs (e.g., *kitten-cat*) the semantic association from the cue to the target words was stronger than the association from the target to the cue words. To illustrate, the likelihood that *kitten* activates *cat* is higher than the likelihood that *cat* activates *kitten*. For the backward-related pairs (e.g., *rain-umbrella*), the association strength had the reverse pattern. Previous research found that backward-related pairs are less likely to be remembered than forward-related ones, but that people do not realize this (Koriat & Bjork, 2005). Following initial studying, a Restudy group studied all pairs again while an Interim Test group undertook a cued recall test, then both groups restudied these pairs in a self-paced procedure. At the restudy phase, the Restudy group spent the same amount of time restudying the forward- and backward-related pairs. In contrast, the Interim Test group spent more time restudying the backward- than forward-related pairs. These findings reveal that interim tests can improve the effectiveness of

self-regulated study time allocation when learning tested information. The question of whether or not interim tests can influence self-regulated learning of new information has not been explored yet. Our Experiments 1 and 2 were designed to investigate whether or not interim tests can modify study time allocation across lists and improve retention of new information.

**Testing effect and metamemory monitoring**

Prior research has found that people tend to be unaware of the backward testing benefit (Agarwal, Karpicke, Kang, Roediger, & McDermott, 2008; Kornell & Son, 2009; Roediger & Karpicke, 2006b). For example, Roediger and Karpicke (2006b) explored the backward testing effect on metamemory monitoring (a form of metacognitive reflection of learning or memory status) by asking participants either to study a text four times or to study it once and take three free recall tests. Then participants estimated what their performance would be in a test to be given one-week later. Participants rated the restudied text more retrievable than the repeatedly tested one, while their test performance showed the reverse pattern.

To date, only one study has employed the multilist paradigm to investigate forward and backward testing effects on metamemory calibration. Szpunar, Jing, and Schacter (2014) divided an online statistics lecture into 4 segments and tested participants either after none of the segments, only after the final one, or after each segment. After the completion of Segment 4, all participants were asked to make a global judgement of learning (JOL; i.e., a way of measuring people's perception of the state of their learning of some material) on the entire lecture to estimate their performance in a final cumulative test. In Szpunar et al.'s study, JOLs overestimated actual recall in the absence of any tests, but four interim tests boosted final recall to the level of predicted recall. In contrast, although a single test did not enhance recall, it did reduce the level of predicted recall. In this study, participants' global JOLs might be affected by both backward as well as forward testing effects. For instance, the interim tests might enhance recall (via a backward effect) which in turn could boost global JOLs. Szpunar et al.'s (2014) research explored the effect of interpolated testing on global JOLs, but the effect of interim testing on list-by-list JOLs has not yet been investigated.

Going beyond Szpunar et al.'s (2014) study, in our Experiments 3 and 4, participants were asked to make a JOL on each separate list to estimate their performance in a possible interim test. By directly measuring changes in JOLs across successive lists, we ask whether people are metacognitively aware of (a) the reduction in retention across successive lists that will occur in the absence of interim tests, as in Szpunar et al.'s (2008) Experiment 2, and (b) the fact that retention will be maintained across lists when interim tests are administered following each list. By measuring the difference in final list JOLs between groups, we are able to ask (c) whether or not JOLs are sensitive to the forward testing effect.

**Present experiments**

In all previous research investigating the forward testing effect, the initial encoding phase was experimenter-paced, which is not common outside the formal classroom. As Schacter and Szpunar (2015) noted, it is important for researchers to assess the extent to which interim testing enhances self-paced learning of new information. Going beyond prior research, in our Experiments 1 and 2, participants were allowed to spend as much time as they wanted to study each item (Euskara-English word pairs or face-name pairs). Self-paced study is of course more typical of real-life learning situations than experimenter-paced study. Moreover, a self-paced procedure enables us to directly explore the possible mechanism underlying the forward testing effect. For example, directly measuring self-regulated study time allocation enables us to shed new light on the contributions of variations in encoding processes to the forward testing effect. In addition, by measuring the amount of PI in the final list interim test, we can determine whether or not variations in retrieval processes constitute another possible source of this effect. In Experiments 3 and 4, the main aim is to explore the forward testing effect on metamemory monitoring and to explore people's metacognitive insight into this forward testing benefit.

Some previous studies have explored the forward testing effect by comparing an Interim Test group, who underwent interim testing following each list, and a No Interim Test group, who performed a distractor task following each list and underwent interim testing following the final list only (Szpunar et al., 2008; Weinstein et al., 2014; Weinstein et al., 2011). Other studies have explored

the forward testing effect by comparing Interim Test, No Interim Test, and Interim Restudy conditions, with the Interim Restudy group restudying after each list and taking an interim test following the final list (Pastötter et al., 2011; Szpunar et al., 2014; Szpunar et al., 2013; Szpunar et al., 2008). All of these latter studies showed that final list interim test recall in the Interim Restudy group was slightly but consistently worse than that in the No Interim Test group. It is reasonable to assume that the Interim Restudy group expected a restudy opportunity when learning the final list, which might have obviated the need to fully encode the final list and impaired its retention (Henkel, 2014; Sparrow, Liu, & Wegner, 2011). Supporting this hypothesis, Sparrow et al. (2011) tested the effect of saving information on a computer. For information that was erased from the computer, participants' recall in a later test significantly outperformed recall of the information saved on the computer, presumably because participants expected that they could re-access the saved information later, which thus reduced the need to fully encode it. Therefore, in the current study, we explored the forward testing effect by comparing Interim Test and No Interim Test conditions.

## Experiment 1

Experiment 1 was conducted to determine how interim tests influence subsequent encoding time allocation when learning new information. In previous research the forward testing effect has been studied only under experimenter-paced conditions. Another aim therefore was to determine whether or not the forward testing effect can be replicated when the encoding procedure is self-paced, which is more typical of self-regulated learning.

### Method

#### Participants

Thirty participants, 24 females, with an average age of 24.10 years (*SD* = 7.22) were recruited from the University College London (UCL) participant pool. Their first language was English. All of them were naïve to the aim of the experiment and reported no prior experience of Euskara, the language of the Basque region. They gave informed consent and reported normal or corrected-to-

normal vision. Participants were randomly divided into two groups (Interim Test/No Interim Test). They were debriefed and received £5 or course credit as compensation after finishing the experiment.

**Materials**

Fifty Euskara nouns with corresponding English translations were selected from a set constructed by Potts and Shanks (2014) (e.g., *sagu – mouse*). These 50 Euskara nouns were divided into five lists of 10 items each, matched for numbers of syllables and letter length. List order was counterbalanced across participants by a Latin square design: three participants in each group studied these five lists in each of five orders.

**Design and procedure**

The experiment involved a 2 (Interim test: Interim Test/No Interim Test) × 5 (List: 1-5) mixed design. Interim test was manipulated between-subjects and List within-subjects. The experiment was conducted in an individual sound-proofed testing room and presented on a computer display using MATLAB software.

Participants were informed that they would study five lists of Euskara-English word pairs in anticipation of a cumulative test. Their task was to commit each Euskara word and its translation to memory. They were also informed that, after studying each list and solving math problems for 1 min, the computer program would randomly decide whether or not to give them a short test. If it did, they would undertake a test of the 10 pairs just studied. If it did not, they would continue solving math problems for another 1.5 min. In fact, participants in the Interim Test group were tested on all five lists while those in the No Interim Test group were only tested on List 5 (see the experimental design schema in Figure 1).

At each list's encoding stage, 10 pairs were presented one at a time in a random order. Participants had unlimited time to study each pair and pressed *ENTER* to end studying the current pair. After studying each individual list, they solved as many math problems (e.g., *47 + 38 = ____?*) as they could in 1 min. Then they continued solving math problems for another 1.5 min or took a short test. At the interim test stage, Euskara cue words from the preceding list were presented in a random

order and participants had unlimited time to recall and type in each word's English translation. Following the completion of List 5, a cumulative recall test was administered. All 50 Euskara words were presented one by one in a random order, and participants had unlimited time to recall each word's translation and type it via the keyboard. There was no feedback in the interim and cumulative tests, and participants were allowed not to respond to a Euskara word if they did not remember its translation.

## Results

### Encoding time

The mean encoding time per word pair on each of Lists 1-5 for both groups is shown in Figure 2A. These data were analysed by a mixed analysis of variance (ANOVA) with Interim test as a between-subjects variable and List (1–5) as a within-subjects variable. A within-subjects contrast showed that there was a negative linear regression of study time across lists, $F(1, 28) = 14.41$, $p < .01$, $\eta_p^2 = .34$, as well as a linear interaction between List and Interim test, $F(1, 28) = 5.63$, $p = .03$, $\eta_p^2 = .17$. Interim test had no main effect, $F(1, 28) = 1.92$, $p = .177$, $\eta_p^2 = .06$. Subsequent repeated-measures ANOVAs, with List as a within-subjects variable, showed that participants in the No Interim Test group decreased their encoding time linearly across lists, $F(1, 14) = 19.73$, $p < .01$, $\eta_p^2 = .59$. In contrast, in the Interim Test group, there was no main effect of List, $F(4, 56) = .74$, $p = .57$, $\eta_p^2 = .05$.

Overall, participants in the No Interim Test group decreased their study time linearly across lists, whereas study time in the Interim Test group did not decline across lists. An independent-samples *t* test revealed that participants in the Interim Test group spent more time encoding List 5 items than participants in the No Interim Test group, mean difference = 4.25 sec per word pair, 95% confidence interval (CI) [.66, 7.84]. There was no significant difference in study time between the groups on any of Lists 1-4, $0.7 \leq t\text{s} \leq 1.85$, $.95 \geq p\text{s} \geq 0.08$.

### Interim test recall and intrusions

Figure 2B shows interim test recall on List 5 for the No Interim Test group and on each of Lists 1-5 for the Interim Test group. Participants in the Interim Test group recalled about 70% of translations across lists, and their recall did not fluctuate across lists, $F(4, 56) = 1.11$, $p = .36$, $\eta_p^2 = .07$. The critical comparison of interim test recall between the groups was on List 5. Levene's test showed that the assumption of homogeneity of variance was not met, $F(1, 28) = 8.38$, $p < .01$. With adjustment, the results showed that participants in the Interim Test group recalled more List 5 translations than participants in the No Interim Test group, mean difference = 2.60 [.83, 4.37] translations.

Even though fewer incorrectly recalled pairs in the Interim Test group than in the No Interim Test group meant fewer opportunities for intrusions (mistakenly recalling another word's translation from any list including the current list) in the List 5 interim test, the overall difference in intrusions between the groups was not statistically significant (No Interim Test group: $M = 2.47$, $SD = 1.46$; Interim Test group: $M = 1.47$, $SD = 1.85$), mean difference = 1.00 [-.24, 2.24] translations. However when the analysis is restricted to intrusions from prior lists, participants in the No Interim Test group experienced more PI in the form of intrusions (mistakenly recalling another word's translation from a prior list) ($M = 1.67$, $SD = 1.35$) than participants in the Interim Test group ($M = .60$, $SD = 1.45$), mean difference = 1.07 [.02, 2.11] translations. No significant difference in intrusions from the current list between the groups was detected (No Interim Test group: $M = .80$, $SD = .77$; Interim Test group: $M = .87$, $SD = 1.06$), mean difference = -.07 [-.76, .63] translations. Of all intrusions, 32.4% were from the current list in the No Interim Test group, compared to 59.2% in the Interim Test group. These results imply that participants in the Interim Test group were better able to control their retrieval from the current list and that the memory search set in the Interim Test group was smaller than that in the No Interim Test group (Weinstein et al., 2011).

**Cumulative test recall**

Overall, participants in the Interim Test group outperformed participants in the No Interim Test group in the cumulative test. We analyse the data separately for List 1-4 pairs and List 5 pairs. Two factors can explain any difference observed in recall of List 1-4 pairs: first, as already

demonstrated, these items were studied for longer in the Interim Test group (the forward testing effect); secondly, they may have benefitted from a backward retrieval practice effect, as these items were tested after each list in one group but not the other. The theoretical analysis of List 5 recall includes two potential factors: first, a forward effect of prior testing; secondly, because the level of recall on the List 5 interim test was higher in the Interim Test group, a greater backward testing effect for the List 5 interim test may have occurred (Rowland, 2014).

As shown in Figure 2C, participants in the Interim Test group recalled more List 1-4 translations than participants in the No Interim Test group, mean difference = 9.81 [2.63, 16.98] translations. This is a very large difference, roughly a doubling of the number of targets recalled. The two factors mentioned above may be contributing. The group difference is evident on List 1, where study time is the same across groups. This can be partially attributed to the fact that testing improves learning and retention, or to the fact that the Interim Test group benefited from additional exposure because they effectively re-experienced those Euskara-English word pairs that they were able to recall in the interim tests. But the effect gets somewhat larger on subsequent lists, suggesting a role for differential encoding time. Participants in the Interim Test group also successfully recalled more List 5 translations than participants in the No Interim Test group, mean difference = 2.07 translations, although this is only marginally significant, 95% CI [-.11, 4.24].

**Correlations between study time and interim test recall**

For each group, we calculated a Pearson correlation between the average study time on List 5 and interim test recall for that list across participants. For both groups there was a positive correlation, but neither of them was statistically significant (No Interim Test group: $r$ = .36, $p$ = .19; Interim Test group: $r$ = .28, $p$ = .31). When collapsed across groups to increase power, the correlation was positive and statistically significant, $r$ = .45, $p$ = .01.

**Discussion**

The results reveal that in the absence of interim tests, participants decreased their encoding time across lists. In contrast, encoding time did not decrease across lists in the Interim Test group.

Thus testing has a forward benefit under conditions of self-paced study and can maintain people's motivation to commit time to studying new information. In line with the decrease in encoding time, participants in the No Interim Test group recalled fewer translations in the List 5 interim test than participants in the Interim Test group. These results provide support for the encoding theory: testing of studied information can directly influence encoding of new information (in this case, measured via study time). In the List 5 interim test, less PI was experienced in the Interim Test group, which provides support for the retrieval theory: this theory proposes that testing has a forward benefit via enriched contextual list information, which differentiates untested information from tested information; PI provides an index of this enhanced list differentiation.

## Experiment 2

To generalize and conceptually replicate the findings of Experiment 1, in Experiment 2 we employed 4 lists of 12 face-name pairs as the experimental materials, as Weinstein et al. (2011) did. This permits us to ask whether or not the forward testing effect on self-regulated study time allocation extends to face-name learning.

### Method

#### Participants

Forty participants, 31 females, with an average age of 23.80 years (*SD* = 5.19) were recruited from the UCL participant pool. Their first language was English. All participants gave informed consent and reported normal or corrected-to-normal vision. They were randomly divided into two groups (Interim Test/No Interim Test). They were debriefed and received £5 or course credit as compensation after finishing the experiment.

#### Materials

Forty-eight male face pictures were collected from the Psychological Image Collection at Stirling (PICS) (available from: http://pics.psych.stir.ac.uk/), the same source used by Weinstein et al. (2011). In addition 48 male names were collected from TOP BABY BOY NAMES 2014, Baby

Centre UK (available from: http://www.babycentre.co.uk/a25011625/top-baby-boy-names-2014#ixzz3TbXFW52d). The faces and names were randomly paired and then were divided into 4 lists of 12 pairs each. Face-name assignments were consistent across participants. List order was counterbalanced by a Latin square design: 5 participants in each group studied these lists in each of 4 orders.

**Design and procedure**

Experiment 2 involved a 2 (Interim test: Interim Test/ No Interim Test) × 4 (List: 1-4) mixed design. As in Experiment 1, Interim test was manipulated between-subjects and List within-subjects. Participants were informed that they would study 4 lists of face-name pairs in anticipation of a cumulative test. Each list consisted of 12 pairs. Faces were presented on the left side and names on the right side of the screen. Participants had unlimited time to study each pair. After studying each individual list, they had 1 min to solve as many math problems as they could. Then, they might or might not be asked to continue solving math problems for another 1.5 min or be asked to take a cued recall test of the 12 pairs just studied. As before, participants were told that the computer program would randomly decide whether or not to give them a short test. In fact, participants in the Interim Test group were tested on every individual list, while participants in the No Interim Test group were only tested on List 4 (see the experiment design schema in Figure 1). Following the completion of List 4, all 48 faces were presented one by one in a random order, and participants had unlimited time to recall each face's name and type it via the keyboard. There was no feedback on the interim and cumulative tests, and participants were allowed not to respond to a face if they did not remember its corresponding name.

**Results**

**Encoding time**

The mean encoding time per face-name pair on each of Lists 1-4 for both groups is shown in Figure 3A. These data were analysed by a mixed ANOVA, with Interim test as a between-subjects variable and List as a within-subjects variable. There was no main effect of List, $F(3, 114) = .48$, $p$

= .69, $\eta_p^2$ = .01, but there was a main effect of Interim test, $F(1, 38) = 7.14$, $p = .01$, $\eta_p^2 = .16$. There was also an interaction between the linear trend of List and Interim test, $F(1, 38) = 12.09$, $p < .01$, $\eta_p^2 = .24$. Repeated-measures ANOVAs showed that participants in the No Interim Test group decreased their study time linearly across lists, $F(1, 19) = 10.33$, $p < .01$, $\eta_p^2 = .35$, whereas participants in the Interim Test group tended to increase their encoding time linearly across lists, $F(1, 19) = 4.36$, $p = .05$, $\eta_p^2 = .19$. Participants in the Interim test group spent more time encoding Lists 2 (mean difference = 2.94 [.27, 5.62] sec per pair), 3 (mean difference = 5.11 [1.51, 8.71] sec), and 4 (mean difference = 8.35 [3.58, 13.12] sec) than those in the No Interim Test group. There was no significant difference in List 1 encoding time between the groups, mean difference = -.81 [-4.20, 2.58] sec.

**Interim test recall and intrusions**

Figure 3B shows interim test recall on List 4 for the No Interim Test group and on each of Lists 1-4 for the Interim Test group. In the Interim Test group, a repeated measures ANOVA, with List as a within-subjects variable, showed that recall tended to increase linearly across lists, $F(1, 19) = 3.40$, $p = .08$, $\eta_p^2 = .15$. Participants in the Interim Test group recalled more List 4 names than participants in the No Interim Test group, mean difference = 3.40 [1.74, 5.06] names, again reflecting a forward testing effect.

Participants in the Interim Test group recalled about 52.9% of names in the List 4 interim test. For the No Interim Test group, only 24.6% were recalled. Even though fewer opportunities were left for intrusions (mistakenly recalling another face's name from any list including the current one) in the Interim Test group than in the No Interim Test group, the difference in overall intrusions between the groups in the List 4 interim test was not statistically significant (No Interim Test group: $M = 5.05$, $SD = 3.10$; Interim Test group: $M = 3.50$, $SD = 2.33$), mean difference = 1.55 [-.21, 3.31] names. Nevertheless participants in the No Interim Test group experienced more PI (mistakenly recalling another face's name from a prior list) ($M = 2.25$, $SD = 1.83$) than participants in the Interim Test group ($M = 1.05$, $SD = 1.36$), mean difference = 1.20 [.17, 2.23] names. The two groups made roughly equivalent numbers of current list intrusions (mistakenly recalling another face's name from the current list, No Interim Test group: $M = 2.80$, $SD = 3.00$; Interim Test group: $M = 2.45$, $SD = 2.21$)

mean difference = .35 [-1.34, 2.04] names. Of all intrusions, 55.5% were from the current list in the No Interim Test group compared to 70.0% in the Interim Test group. These results, replicating those in Experiment 1, indicate that the memory search set in the No Interim Test group was bigger than that in the Interim Test group.

**Cumulative test recall**

As illustrated in Figure 3C, participants in the Interim Test group recalled more List 1–3 names in the cumulative test than participants in the No Interim Test group, mean difference = 9.00 [5.27, 12.73] names. More importantly, participants in the Interim Test group recalled more List 4 names than participants in the No Interim Test group, mean difference = 3.15 [1.46, 4.84] names.

**Correlations between study time and interim test recall**

At the participant level, Pearson correlations between List 4 average study time and interim test recall for that list for each group were not statistically significant (Interim Test group, $r$ = -.03, $p$ = .89; No Interim Test group, $r$ = .21, $p$ = .37). Combining the data across groups to increase power, the correlation was positive and marginally significant, $r$ = .30, $p$ = .06. Although not reaching the conventional level of statistical significance, this is a medium-sized correlation.

**Discussion**

Consistent with Experiment 1, participants in the No Interim Test group in Experiment 2 decreased their encoding time linearly across lists. Participants in the Interim Test group actually increased their encoding time linearly across lists. Thus interim testing boosts self-regulated study time allocation when learning new information. In the List 4 interim test, participants in the Interim Test group successfully recalled more names than participants in the No Interim Test group, while the latter group experienced more PI in the List 4 interim test. Again, Experiment 2's results provide support for both encoding and retrieval factors playing roles in the forward testing effect.

Soderstrom and Bjork (2014) found that interim testing facilitates people's self-regulated study time allocation by alleviating metacognitive unawareness of the difference in recall difficulty

between forward- and backward-related pairs. Our Experiments 1 and 2 found that interim testing facilitates people's self-regulated study time allocation by preventing encoding time reduction across lists. Combining these findings, we conclude that interim testing facilitates self-regulated study time allocation for the encoding of both studied and new information.

## Experiment 3

Experiments 1 and 2, together with prior demonstrations of the forward testing effect (Cho et al., 2016; Pastötter et al., 2011; Szpunar et al., 2013; Szpunar et al., 2008; Weinstein et al., 2014; Weinstein et al., 2011), strongly suggest that learning of new information can be considerably boosted by testing of prior information. In the classroom this benefit can be achieved by the instructor choosing to insert tests during a lesson. However, recent survey results show that learners themselves are reluctant to administer tests during learning (Karpicke, Butler, & Roediger, 2009). Although they may do so in some situations, Kornell and Son (2009) found that people's motivation for self-testing is largely derived from a desire to diagnose their current level of learning, rather than from metacognitive awareness of the enhancing backward effect of testing. Similarly, in the context of self-regulated learning, learners may be less likely to administer interim tests during learning if they lack metacognitive awareness of the forward benefit of testing. In contrast, if they appreciate the forward benefits of interim testing, their motivation to self-administer interim tests may be boosted.

This alignment between an objective benefit on the one hand and metacognitive awareness on the other cannot be taken for granted. Prior research has shown that while testing enhances retention of tested information, people's metamemory indicates that they rate restudying more effective than testing (Roediger & Karpicke, 2006b). The primary aim of Experiment 3 was to explore people's metacognitive insight regarding this forward testing benefit through measuring their list-by-list judgements of learning. Face-name pairs were again used. The key aim was to determine whether people's JOLs are sensitive to the reduction of learning across lists in the absence of interim tests, and whether their JOLs are sensitive to the fact that their retention will be maintained across lists when interim tests are administered following each list.

In the previous two experiments, participants in the No Interim Test group decreased their encoding time across lists. In contrast to the previous two experiments, in Experiment 3, we used an experimenter-paced procedure. The main reason for this is that participants' JOLs can be directly affected by their study time allocation. For example, in a self-paced condition, the No Interim Test group will substantially decrease their encoding time across lists as suggested by our Experiments 1 and 2, and the reduction of encoding time in turn may directly decrease JOLs across lists. To remove the direct influence of study time allocation on JOLs, here we employed the experimenter-paced procedure, which enables us to explore whether or not participants in the No Interim Test group can appreciate the decrease of their learning effectiveness and whether or not participants in the Interim Test group can appreciate the maintenance of their learning effectiveness across lists when the encoding phase is experimenter-paced (Pastötter et al., 2011; Szpunar et al., 2013; Wissman et al., 2011). Another reason is that previous studies showed that asking participants to make JOLs affects their self-regulated study time allocation. For example, when expecting to make JOLs people may spend some time considering the memorability of an item, and devote a portion of the encoding time to assessing their on-going learning status (Mitchum, Kelley, & Fox, 2016). Therefore, to directly explore the forward testing effect on metamemory monitoring, we employed an experimenter-paced procedure.

## Method

### Participants

Forty participants, 30 females, with an average age of 23.13 years (*SD* = 5.25) were recruited from the UCL participant pool and randomly divided into two groups (Interim Test/No Interim Test). Their first language was English. They gave informed consent and reported normal or corrected-to-normal vision. After finishing the experiment, they were debriefed and received £4 or course credit as compensation.

### Materials, design, and procedure

The same materials, design, and procedure were used as in Experiment 2 with the following exceptions. During each list's encoding phase, participants had 4 sec to study each pair, as Weinstein et al. (2011) did. After studying each individual list, participants were asked to make a JOL. They estimated how many names they thought they would be able to recall correctly if they were tested on the 12 just-studied pairs in 1 min. JOLs were made on a slider ranging from 0 ("I won't recall any names correctly") to 12 ("I will recall all names correctly") (see the experiment design schema in Figure 1).

**Results**

**JOLs**

Average JOLs on each of Lists 1-4 for both groups are shown in Figure 4A. These data were analysed by a mixed ANOVA, with Interim test as a between-subjects variable and List as a within-subjects variable. Tests of within-subjects contrasts showed that JOLs decreased linearly across lists, $F(1,38) = 23.17$, $p < .01$, $\eta_p^2 = .38$, and there was a linear interaction between Interim test and List, $F(1,38) = 5.23$, $p = .03$, $\eta_p^2 = .12$. Interim test had no main effect, $F(1, 38) = 2.85$, $p = .10$, $\eta_p^2 = .07$. For the No Interim Test group, a follow-up repeated-measures ANOVA with List as a within-subjects variable showed that there was a negative linear regression of JOLs across lists, $F(1, 19) = 28.52$, $p < .01$, $\eta_p^2 = .60$. For the Interim Test group, a similar ANOVA revealed no main effect of List, $F(3, 57) = 1.14$, $p = .34$, $\eta_p^2 = .06$.

The linear interaction between Interim test and List indicates that the No Interim Test group decreased their JOLs across lists more than the Interim Test group. Specifically, participants in the Interim Test group gave higher JOLs than participants in the No Interim Test group on Lists 3 (mean difference = 1.20 [.15, 2.25] names) and 4 (mean difference = 1.15 [.27, 2.03] names). No statistically significant difference between the two groups' JOLs on Lists 1 and 2 was detected, $.19 \leq t \leq .45$, $.65 \leq p \leq .85$.

**Interim test recall and intrusions**

Interim test recall on List 4 for the No Interim Test group and on each of Lists 1-4 for the Interim Test group is shown in Figure 4B. For the Interim Test group, the data were analysed by a repeated-measures ANOVA with List a within-subjects variable. The assumption of sphericity was not met, $\chi^2(5) = 16.73$, $p < .01$, so we applied the Huynh-Feldt correction. The ANOVA revealed no main effect of List, $F(2.13, 40.38) = .02$, $p = .998$, $\eta_p^2 = .001$, indicating that participants' interim test recall did not vary systematically across lists. In the List 4 interim test, participants in the Interim Test group recalled more names than participants in the No Interim Test group, mean difference = 2.00 [.53, 3.47] names.

Although participants in the No Interim Test group generated numerically more intrusions (mistakenly recalling another face's name from any list including the current one) in the List 4 interim test ($M = 5.75$, $SD = 3.66$) than participants in the Interim Test group ($M = 4.30$, $SD = 3.28$), the difference between the groups was not statistically significant, mean difference = 1.45 [-.78, 3.68] names. To measure PI (mistakenly recalling another face's name from a prior list), we ran an independent-samples *t* test. Levene's Test revealed inequality of variances, $F(1, 38) = 9.56$, $p < .01$. With adjustment, the results showed that participants in the No Interim Test group experienced more PI ($M = 2.90$, $SD = 2.22$) than those in the Interim Test group ($M = .85$, $SD = 1.23$), mean difference = 2.05 [.90, 3.20] names. No significant difference in current list intrusions was detected between the groups (No Interim Test group: $M = 2.85$, $SD = 2.30$; Interim Test group: $M = 3.45$, $SD = 3.20$; mean difference = -.60 [-2.39, 1.19] names. Of all intrusions, 49.6% were from the current list in the No Interim Test group, far fewer than in the Interim Test group (80.2%). These results indicate once again that the memory search set was larger in the No Interim Test than in the Interim Test group.

**Cumulative test recall**

In the cumulative test, participants in the Interim Test group recalled more List 1-3 names than participants in the No Interim Test group, mean difference = 4.28 [1.94, 7.76] names (see Figure 4C). This can be attributed to the fact that testing improves retention, and to the fact that more attention and effort might be directed to learning Lists 2-3 (Pastötter et al., 2011; Szpunar et al.,

2013). In addition, participants in the Interim Test group recalled more List 4 names than participants in the No Interim Test group, mean difference = 2.55 [1.16, 3.94] names.

Appendix A reports JOL calibration (absolute agreement between judgements of learning and recall; see detailed explanation in Appendix A.) and correlation results. There was no difference in List 4 JOL calibration between the groups. Interestingly, the Interim Test group showed a greater correlation between List 4 JOLs and List 4 interim test recall than the No Interim Test group (see details in Appendix A).

**Discussion**

Participants in the No Interim Test group reduced their JOLs across lists much more than those in the Interim Test group. Importantly, List 4 JOLs were aligned with List 4 interim test recall: both recall and JOLs were significantly higher in the Interim Test group than in the No Interim test group, revealing that both retention and metamemory monitoring are influenced in a similar way by the effect of prior interim tests. The same pattern in List 4 JOLs and interim test recall (higher JOLs and recall in the Interim Test group than in the No Interim Test group) reveals participants' metacognitive insight into the forward testing benefit. The Interim Test group suffered less PI in the List 4 interim test than the No Interim Test group, which again supports the involvement of retrieval processes in the forward testing effect.

We replicated the forward testing effect when the procedure was experimenter-paced. Therefore, this effect cannot be simply attributed to the additional exposure time (more encoding time of the final list in the Interim Test group than in the No Interim Test group) that is available when study is self-paced (as in Experiments 1 and 2). The possible mechanisms underlying the forward testing effect in self- and experimenter-paced conditions are further discussed later.

## Experiment 4

To generalize and conceptually replicate the findings of Experiment 3, in Experiment 4 we employed 5 18-word lists as materials, as Szpunar et al. (2008) did. Thus the materials were single words rather than foreign language translations or face-name pairs.

## Method

### Participants

Forty participants, 36 females, with an average age of 19.70 years (*SD* = 3.64) were recruited from the UCL participant pool and randomly divided into two groups (Interim Test/No Interim Test). Their first language was English. They gave informed consent and reported normal or corrected-to-normal vision. After finishing the experiment, they were debriefed and received £4 or course credit as compensation.

### Materials

Ninety English nouns were drawn from the MRC Psycholinguistic Database (available from: http://websites.psychology.uwa.edu.au/school/MRCDatabase/uwa_mrc.htm). Letter length was controlled between 4 and 8, Kucera-Francis written frequency between 100 and 850, and concreteness and familiarity between 250 and 650. These nouns were randomly divided into 5 lists of 18 items each. List order was counterbalanced across participants by a Latin square design: 4 participants in each group studied these lists in each of 5 orders.

### Design and procedure

Experiment 4 involved a 2 (Interim test: Interim Test/No Interim Test) × 5 (List: 1-5) mixed design. Interim test was a between-subjects variable and List was a within-subjects variable. The procedure was similar to that of previous experiments, except as noted. Participants were instructed to study 5 lists of English words and were warned that a cumulative free recall test would be administered following the completion of List 5. They were informed that after encoding each list the computer would decide at random whether or not to give them a short test. In fact, the No Interim Test group was only tested on List 5 and the Interim Test group was tested on every list (see the experiment design schema in Figure 1).

At the encoding stage, each word was presented for 2 sec for participants to study, as Szpunar et al. (2008) did. After studying each individual list, participants predicted what proportion of words

from that list they thought they would be able to recall if they were tested in 1 min. JOLs were made on a slider ranging from 0 ("I won't recall any words") to 100 ("I will recall all words"). After that, they solved as many math problems as they could in the next 1 min. Then they undertook a 1 min free recall test or continued solving math problems for another 1 min. After the completion of List 5, participants were asked to freely recall as many words as they could from all 5 lists.

**Results**

**JOLs**

Average JOLs on each of Lists 1-5 for both groups are shown in Figure 5A. These data were analysed by a mixed ANOVA, with Interim test as a between-subjects variable and List as a within-subjects variable. Tests of within-subjects contrasts revealed a negative linear regression of JOLs across lists, $F(1, 38) = 43.66$, $p < .01$, $\eta_p^2 = .54$, and a linear interaction between List and Interim test, $F(1, 38) = 10.88$, $p < .01$, $\eta_p^2 = .22$. Interim test had no main effect, $F(1, 38) = .49$, $p = .49$, $\eta_p^2 = .01$. Participants in both groups decreased their JOLs linearly across lists: No Interim Test group, $F(1, 19) = 37.60$, $p < .01$, $\eta_p^2 = .66$; Interim Test group, $F(1, 19) = 7.88$, $p = .01$, $\eta_p^2 = .29$.

The interaction between Interim test and List reveals that participants in the No Interim Test group decreased their JOLs across lists more than participants in the Interim Test group. Specifically, participants in the Interim Test group gave higher JOLs on List 5 than participants in the No Interim Test group, mean difference = 11.8% [.18, 23.42]. No statistically significant difference in JOLs was detected on Lists 1-4, $-1.57 \leq t \leq .68$, $.13 \leq p \leq .50$.

**Interim test recall and intrusions**

Interim test recall on List 5 for the No Interim Test group and on each of Lists 1-5 for the Interim Test group is shown in Figure 5B. For the Interim Test group, a repeated measures ANOVA with List as a within-subjects variable showed that there was no main effect of List, $F(4, 76) = .24$, $p = .92$, $\eta_p^2 = .01$, indicating that participants' interim test recall did not vary systematically across lists. In the List 5 interim test, participants in the Interim Test group recalled more List 5 words than participants in the No Interim Test group, mean difference = 4.30 [2.38, 6.22] words. In this

experiment intrusions can only be from prior lists; current list intrusions are not meaningful because the test was free recall. Participants in the No Interim Test group ($M$ = 3.30, $SD$ = 4.27) experienced substantially more PI (intrusions from prior lists) in the List 5 interim test than participants in the Interim Test group ($M$ = .25, $SD$ = .55), mean difference = 3.05 [1.10, 5.00] words.

**Cumulative test recall**

In the cumulative test, participants in the Interim Test group recalled more List 1-4 words than participants in the No Interim Test group, mean difference = 5.80 words, which is marginally significant, 95% CI [-.02, 11.62] (see Figure 5C). More importantly, participants in the Interim Test group also recalled more List 5 words than participants in the No Interim Test group, mean difference = 2.55 [.47, 4.63] words.

Appendix B reports JOL calibration and correlation results for this experiment. Again, there was no significant difference in List 5 JOL calibration between the groups. Although there was no statistically significant difference in correlations between List 5 JOLs and interim test recall, the Interim Test group showed numerically (albeit not significantly) greater correlation than the No Interim Test group, which is a similar pattern to that in Experiment 3.

**Discussion**

Once again, JOLs in the No Interim Test group decreased across lists much more than those in the Interim Test group, indicating participants' realization that their learning was becoming less effective across lists. JOLs on the final list were aligned with List 5 interim test recall. Less PI was experienced in the Interim Test group than in the No Interim Test group, which again supports the retrieval account of the forward testing effect. Consistent with Experiment 3, we found a forward testing effect when the procedure was experimenter-paced, again supporting the claim that factors in addition to extra exposure time (seen in self-paced conditions) play a role.

## General discussion

In the current research, we first explored the forward testing effect in self-paced conditions in Experiments 1 and 2. Consistent with previous studies (Pastötter et al., 2011; Szpunar et al., 2013; Weinstein et al., 2011), a strong forward testing effect was obtained. People may be reluctant to administer interim tests during their learning if they are unaware of this forward testing benefit. In other words, metacognitive awareness of this forward testing effect may boost the likelihood of self-administering interim tests. Next, in our Experiments 3 and 4 we explored whether people tend to be aware of the forward testing benefit through measuring list-by-list JOLs. Across all four experiments, the forward testing effect was replicated no matter whether the encoding procedure was self- or experimenter-paced. In the final list interim test, participants in the Interim Test group outperformed those in the No Interim Test group. This effect was substantial and amounted to an approximate doubling of final list interim test recall. In Experiments 1 and 2, we observed a decreasing slope of encoding time across lists in the No Interim Test group which was not present in the Interim Test group. Indeed in Experiment 2, the Interim Test group's encoding time increased across lists. Thus, as indexed by self-controlled study time, the preceding tests served to maintain motivation to engage in effective encoding. In all experiments, we saw evidence that this forward benefit of interim tests was associated with a reduction in the amount of proactive interference experienced in the final list interim test. In Experiments 3 and 4, participants' JOLs decreased across lists in both groups, but JOLs in the Interim Test group decreased much less across lists than those in the No Interim Test group.

In Experiments 1 and 2, participants in the Interim Test group spent more time encoding the final list than participants in the No Interim Test group, which supports the claim that variations in encoding processes (e.g., attention) play a role in the forward testing effect (Pastötter et al., 2011; Szpunar et al., 2013; Wissman et al., 2011). The difference in interim test recall of the final list (the forward testing effect) can be partially attributed to variations in the encoding process, as indicated by the difference in encoding time between groups. The Interim Test group committed more time to encoding the final list than the No Interim Test group, and more encoding time produces superior learning and memory (as indicated by the positive correlation between encoding time and interim test recall in Experiments 1 and 2). Although the learning procedure was experimenter-paced in

Experiments 3 and 4, the difference in interim test recall of the final list can also be partially attributed to variations in encoding processes. Learning should have deteriorated across lists in the No Interim Test group and remained constant in the Interim Test group even when the learning procedure was experimenter-paced. Pastötter et al. (2011) found patterns of constancy (Interim Test group) and increase (No Interim Test group) in alpha power – an index of reduced attention – across lists when the learning procedure was experimenter-paced. Similarly, Jing et al. (2016) found that an Interim Test group reported fewer task-unrelated thoughts than a No Interim Test group when the learning procedure was experimenter paced. Besides variations in encoding processes, the release from PI in the Interim Test group observed in all four experiments supports the alternative but not mutually exclusive idea that facilitation of retrieval is partly responsible for the forward testing effect in both self- and experimenter-paced situations (Szpunar et al., 2008). In Experiments 1-3, in the final list interim test, higher proportions of intrusions were from the current list in the Interim Test group than in the No Interim Test group, indicating that the memory search set in the Interim Test group was smaller and that these participants were better able to control their recall from the current list (Bäuml & Kliegl, 2013; Weinstein et al., 2011), which again supports the retrieval account.

Why exactly do interim tests protect against the decrease of encoding time across lists that is observed in the absence of interim tests? Prior research has found that test expectancy (knowing that one will be tested) plays an important role in encoding and long-term retention (Nestojko, Bui, Kornell, & Bjork, 2014; Szpunar, McDermott, & Roediger, 2007; Weinstein et al., 2014). Specifically, the effect of interim tests may be mediated by test expectancy, which in turn boosts learning motivation. Weinstein et al. (2014) employed a multiple list procedure to investigate the *test expectancy effect* on release from PI. Participants' test expectancy in the Interim Test group remained fairly constant across lists. However, test expectancy in the No Interim Test group decreased – perhaps unsurprisingly – across lists. In the final list interim test, a forward testing effect was observed and interim tests alleviated PI, as found here. Thus the forward testing effect may, at least in part, be attributed to the fact that interim tests act as warnings of the upcoming test, which forces people to focus their attention and effort on encoding new information. In our Experiments 1 and 2,

participants in the No Interim Test group presumably decreased their test expectancy across lists and accordingly decreased their encoding time across lists. Of course, both groups knew there would be a final cumulative test, but the immediacy of the interim tests was presumably more effective than the prospect of a more remote cumulative test in maintaining motivation.

Pastötter et al. (2011) proposed an *encoding reset theory* to account for why interim testing prevents deterioration of subsequent encoding of new information. Pastötter, Bäuml, and Hanslmayr (2008) manipulated participants' mental context between-subjects when studying two lists of words and recorded brain activity during encoding. In a context change condition, after studying the first list, participants were instructed to imagine walking through their parents' house and describe their mental imagery, which induced an internal context change. In a control condition, participants did not perform the imagination task. The researchers observed superior recall of the second list in the context change compared to the control condition. Correspondingly, they found that theta and alpha power, which are linked to reduced attention, increased from the first to the second lists in the control but not the context change condition. These findings suggest that mental context change induces a 'reset' of encoding of the second list, making encoding of the second list as effective as encoding of the first one. Further evidence comes from research by Pastötter et al. (2011). Pastötter et al. (2011) found a significant increase of alpha power across lists in a No Interim Test group, but no such increase in an Interim Test group, suggesting that interim testing induces an internal context change between lists which induces a reset of encoding of the subsequent list, making its encoding as effective as that of the prior lists.

Why did interim tests lead to an increase of encoding time across lists when using face-name pairs in Experiment 2 but not when using Euskara-English pairs in Experiment 1? Prior research has found that people overestimate their learning when encoding is fluent (Hertzog, Dunlosky, Robinson, & Kidder, 2003). Face-name encoding is common in daily life, whereas in Experiment 1 no participant reported any prior study experience of Euskara. It seems reasonable therefore to speculate that face-name encoding is more fluent than Euskara-English encoding. If, on the basis of their experienced fluency, participants overestimated their learning of face-name pairs in List 1 relative to

the situation with word pairs, then the interim tests might have served to calibrate their assessments of learning and made them realize the gap between their perceived and actual learning status. In Experiment 3, even when the encoding procedure was experimenter-paced and encoding time was shorter than that in Experiment 2, participants in the Interim Test group overestimated their face-name learning on List 1 (JOLs: 4.90 names; Interim test recall: 4.40 names) and then the List 1 interim test calibrated their List 2 JOLs (JOLs: 4.40 names; Interim test recall: 4.30 names). A prediction of this account is that in the first list, the gap between JOLs and recall might be greater for face-name than Euskara-English pairs, something not evaluated in the present experiments. Another possible reason is that faces and names were subjectively more similar across lists in Experiment 2 than Euskara and English words in Experiment 1. Participants might worry about PI much more when learning face-name pairs than when learning Euskara-English word pairs. Therefore, they increased encoding time when encoding face-name pairs but not when encoding word pairs in Experiment 1. Future research should be conducted to further investigate why intervening tests have different effects on encoding time for different types of materials.

We have interpreted the results as providing some support for the idea that intervening tests facilitate retrieval processes by reducing PI (Szpunar et al., 2008; Weinstein et al., 2011). In Experiment 4, in the final list interim test, participants in the No Interim Test group experienced about 13.20 times more PI (intrusions from preceding lists) than participants in the Interim Test group. However, in Experiments 1-3, in the final list interim tests, participants in the No Interim Test groups suffered only about 2.78, 2.14, and 3.41 times more PI as in the Interim Test group. Why might this substantial difference have occurred? Interim tests generate greater list discrimination by enriching list-specific context, which helps people to limit their memory search set and protect their recall from PI. In the free recall test in Experiment 4, list-specific cues are assumed to play an important role in protecting recall from PI – hence the large effect of intervening tests on PI. But the contribution of list-specific cues in the cued-recall test in Experiments 1-3 was presumably weaker because participants might rely on the cue-to-target associations – hence a more modest effect of intervening tests on PI (Cho et al., 2016).

Prior research has found that people tend to be unaware of the backward testing benefit (Roediger & Karpicke, 2006b). The present experiments reveal that people's final list JOLs are aligned with final list interim recall. It is possible that, in the No Interim Test group, participants appreciated they would suffer interference from prior lists, and therefore decreased their JOLs across lists. Alternatively, they might try to replay their learning process when they made their JOLs and realize that their minds had wandered more and more across lists (Szpunar et al., 2013) and that they made less and less encoding effort (as found in our Experiments 1 and 2; Pastötter et al., 2011) across lists. Participants in the Interim Test group also decreased their JOLs across lists in Experiments 3 and 4. Specifically, JOLs in the Interim Test group fell from List 1 to List 2, and remained stable or decreased marginally across subsequent lists. We interpret this as indirect evidence that effort and attention in the Interim Test group did not fluctuate across lists (Pastötter et al., 2011; Szpunar et al., 2013). Another possible explanation is that the Interim Test group made subsequent list JOLs according to previous lists' interim test recall. The maintenance of interim test recall across lists informs the Interim Test group of the consistency of their learning across lists. Experiments 3 and 4 showed that people's JOLs are sensitive to the forward testing effect as reflected by the alignment between the final list JOLs and final list interim test recall. Both the Interim Test and No Interim Test groups predicted they would remember about half of the List 1 items if they were tested on List 1. In the No Interim Test group, List 1 JOLs might act as an anchor, and participants decreased their JOLs across lists, yielding final list JOLs that were lower than those in the Interim Test group. Future research might explore whether final list JOLs are aligned with final list interim test recall when no prior list JOLs are made.

Metacognitive insight into the forward testing benefit might be explicit: learners might appreciate that their learning and recall is enhanced *because* they took an earlier test. For example, the Interim Test group might have explicitly experienced the forward testing effect and come to believe that interim testing makes subsequent segment encoding as effective as the encoding of previous segments. In other words, the Interim Test group might explicitly know that interim testing enhances their subsequent learning of new information. This metacognitive insight might, on the other hand, be

implicit. It is possible that prior interim tests maintained the Interim Test group's effort in encoding subsequent new information, and more effort may then have led to greater JOLs compared to those in the No Interim Test group. Therefore, the Interim Test group may have reported higher final list JOLs because they allocated more effort to encoding the final list (and were aware of this), without them knowing explicitly that the reason they allocated more effort was because of the prior interim tests. Put differently, the Interim Test group might not explicitly know that interim testing facilitates subsequent learning. The key differential prediction that these two forms of metacognition make – and that could profitably be explored in future research – is that it is only on the basis of explicit knowledge that learners would actively self-administer tests.

In the final cumulative test, across all four experiments, the Interim Test group significantly outperformed the No Interim Test group. The superior cumulative performance in the Interim Test group constitutes a backward testing effect (Roediger & Karpicke, 2006a, 2006b; Weinstein et al., 2011) because items (except final list items) were initially tested in the Interim Test group but not in the No Interim Test group. However, the lack of a restudy comparison group means we cannot rule out the possibility that the additional exposure to studied information that occurred as a result of interim testing was responsible for the enhanced recall in the cumulative test observed in the Interim Test group, rather than any processes specific to testing. The superior cumulative recall in the Interim Test group can also be partially attributed to the fact that more study time, effort, and attention was directed to the encoding process (our Experiments 1 & 2; Pastötter et al., 2011; Szpunar et al., 2013).

## Implications

Self-regulated learning is increasingly taking place outside as much as inside the formal classroom. How to enhance self-regulated learning is a key concern for learners, educators, and researchers. In our first two experiments, we found that interpolated testing maintains people's motivation to commit study time to encoding new information, which enhances learning and retention. Moreover, Experiments 3 and 4 confirmed previous research showing that interpolated testing is also beneficial for experimenter- or educator-paced learning. These findings justify the

recommendation that learners and instructors should consider administering tests during learning in both self- and educator-paced study situations.

In daily life, learners must often master a large body of information, which can be divided into multiple segments. How to prevent proactive interference is another key concern for learners, educators, and researchers. Across all 4 experiments, our data showed that interpolated testing can prevent intrusions from prior learning segments no matter whether the testing format is cued or free recall, and regardless of whether learning is self- or educator-paced. These findings suggest that learners and educators should administer tests during learning to limit the detrimental build-up of proactive interference.

In the formal classroom, educators may insert interim tests during a lecture and obtain a forward testing benefit. Outside the formal classroom, learners' willingness to self-administer interim tests may be boosted by metacognitive insight into the forward testing benefit. Our Experiments 3 and 4 reveal that people are sensitive to the deterioration of learning across segments in the absence of interim tests and appreciate the maintenance of learning effectiveness when interim tests are administered following each segment. People's metacognitive insight regarding the forward testing benefit may thus encourage self-administration of interim tests.

## Conclusion

Interim testing enhances subsequent encoding of new information and prevents a decrease in encoding time across lists. In addition, interim tests insulate against the build-up of PI. The forward benefits of testing are attributable to both encoding (e.g., greater effort and deeper encoding) and retrieval (e.g., greater list discrimination) processes. The forward testing benefit is associated with metacognitive insight. This study leads to a strong recommendation that interim tests can be profitably used to promote learning of new information whenever learning is self- or instructor-paced.

## References

Abbott, E. E. (1909). On the analysis of the factor of recall in the learning process. *The Psychological Review: Monograph Supplements, 11*(1), 159-177. doi: 10.1037/h0093018

Agarwal, P. K., Karpicke, J. D., Kang, S. H. K., Roediger, H. L., & McDermott, K. B. (2008). Examining the testing effect with open- and closed-book tests. *Applied Cognitive Psychology, 22*(7), 861-876. doi: 10.1002/acp.1391

Arnold, K. M., & McDermott, K. B. (2013). Test-potentiated learning: distinguishing between direct and indirect effects of tests. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 39*(3), 940-945. doi: 10.1037/a0029199

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. doi: 10.1016/j.jml.2012.07.006

Bjork, R. A., Dunlosky, J., & Kornell, N. (2013). Self-regulated learning: beliefs, techniques, and illusions. *Annual Review of Psychology, 64*, 417-444. doi: 10.1146/annurev-psych-113011-143823

Carpenter, S. K. (2009). Cue strength as a moderator of the testing effect: The benefits of elaborative retrieval. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 35*(6), 1563-1569. doi: 10.1037/a0017021

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

Davis, S. D., & Chan, J. C. (2015). Studying on borrowed time: How does testing impair new learning? *Journal of Experimental Psychology: Learning, Memory, and Cognition, 41*(6), 1741-1754. doi: 10.1037/xlm0000126

Finn, B., & Roediger, H. L., 3rd. (2013). Interfering effects of retrieval in learning new information. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 39*(6), 1665-1681. doi: 10.1037/a0032377

Henkel, L. A. (2014). Point-and-shoot memories: The influence of taking photos on memory for a museum tour. *Psychological Science, 25*(2), 396-402. doi: 10.1177/0956797613504438

Hertzog, C., Dunlosky, J., Robinson, A. E., & Kidder, D. P. (2003). Encoding fluency is a cue used for judgments about learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 29*(1), 22-34. doi: 10.1037/0278-7393.29.1.22

Izawa, C. (1969). Comparison of reinforcement and test trials in paired-associate learning. *Journal of Experimental Psychology, 81*(3), 600-603. doi: 10.1037/h0027905

Jing, H. G., Szpunar, K. K., & Schacter, D. L. (2016). Interpolated testing influences focused attention and improves integration of information during a video-recorded lecture. *Journal of experimental Psychology: Applied, 22*(3), 305-318. doi: 10.1037/a0019902.supp

Karpicke, J. D. (2009). Metacognitive control and strategy selection: Deciding to practice retrieval during learning. *Journal of Experimental Psychology: General, 138*(4), 469-486. doi: 10.1037/a0017341

Karpicke, J. D., Butler, A. C., & Roediger, H. L., 3rd. (2009). Metacognitive strategies in student learning: Do students practise retrieval when they study on their own? *Memory, 17*(4), 471-479. doi: 10.1080/09658210802647009

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

Koriat, A., & Bjork, R. A. (2005). Illusions of competence in monitoring one's knowledge during study. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 31*(2), 187 - 194. doi: 10.1037/0278-7393.31.2.187

Kornell, N., & Bjork, R. A. (2008). Optimising self-regulated study: The benefits—and costs—of dropping flashcards. *Memory, 16*(2), 125-136. doi: 10.1080/09658210701763899

Kornell, N., & Metcalfe, J. (2006). Study efficacy and the region of proximal learning framework. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 32*(3), 609-622. doi: 10.1037/0278-7393.32.3.609

Kornell, N., & Son, L. K. (2009). Learners' choices and beliefs about self-testing. *Memory, 17*(5), 493-501. doi: 10.1080/09658210902832915

Metcalfe, J., & Kornell, N. (2005). A region of proximal learning model of study time allocation. *Journal of Memory and Language, 52*(4), 463-477. doi: 10.1016/j.jml.2004.12.001

Mitchum, A. L., Kelley, C. M., & Fox, M. C. (2016). When asking the question changes the ultimate answer: Metamemory judgments change memory. *Journal of Experimental Psychology: General, 145*(2). doi: 10.1037/a0039923

Nestojko, J. F., Bui, D. C., Kornell, N., & Bjork, E. L. (2014). Expecting to teach enhances learning and organization of knowledge in free recall of text passages. *Memory & Cognition, 42*(7), 1038-1048. doi: 10.3758/s13421-014-0416-z

Palva, S., & Palva, J. M. (2007). New vistas for alpha-frequency band oscillations. *Trends in Neurosciences, 30*(4), 150-158. doi: 10.1016/j.tins.2007.02.001

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

Pastötter, B., Bäuml, K. H., & 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. doi: 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. doi: 10.1037/a0021801

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

Potts, R., & Shanks, D. R. (2014). The benefit of generating errors during learning. *Journal of Experimental Psychology: General, 143*(2), 644-667. doi: 10.1037/a0033194

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

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

Roediger, H. L., & Karpicke, J. D. (2006a). The power of testing memory: Basic research and implications for educational practice. *Perspectives on Psychological Science, 17*(3), 249-255. doi: 10.1111/j.1745-6916.2006.00012.x

Roediger, H. L., & Karpicke, J. D. (2006b). Test-enhanced learning: Taking memory tests improves long-term retention. *Psychological science, 17*(3), 249-255. doi: 10.1111/j.1467-9280.2006.01693.x

Roediger, H. L., Putnam, A. L., & Smith, M. A. (2011). Ten benefits of testing and their applications to educational practice. *Psychology of Learning and Motivation-Advances in Research and Theory, 55*, 1-36. doi: 10.1016/B978-0-12-387691-1.00001-6

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

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

Silver, N. C., & Dunlap, W. P. (1987). Averaging correlation coefficients: Should Fisher's z transformation be used? *Journal of Applied Psychology, 72*(1), 146-148. doi: 10.1037/0021-9010.72.1.146

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

Sparrow, B., Liu, J., & Wegner, D. M. (2011). Google Effects on Memory: Cognitive Consequences of Having Information at Our Fingertips. *Science, 333*(6043), 776-778. doi: 10.1126/science.1207745

Szpunar, K. K., Jing, H. G., & Schacter, D. L. (2014). Overcoming overconfidence in learning from video-recorded lectures: Implications of interpolated testing for online education. *Journal of Applied Research in Memory and Cognition, 3*(3), 161-164. doi: 10.1016/j.jarmac.2014.02.001

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

Szpunar, K. K., McDermott, K. B., & Roediger, H. L. (2007). Expectation of a final cumulative test enhances long-term retention. *Memory & Cognition, 35*(5), 1007-1013. doi:10.3758/BF03193473

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

Weinstein, Y., Gilmore, A. W., Szpunar, K. K., & McDermott, K. B. (2014). The role of test expectancy in the build-up of proactive interference in long-term memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 40*(4), 1039-1048. doi: 10.1037/a0036164.supp

Weinstein, Y., McDermott, K. B., & Szpunar, K. K. (2011). Testing protects against proactive interference in face-name learning. *Psychonomic bulletin & review, 18*(3), 518-523. doi: 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. *Psychonomic bulletin & review, 18*(6), 1140-1147. doi: 10.3758/s13423-011-0140-7

## Footnote

1. Several terms have been used to refer to the fact that testing can enhance learning and retention of new information: *the interim test effect* (Wissman et al., 2011), *the facilitative effect of interpolated testing on subsequent learning* (Szpunar et al., 2013), *test-enhanced new learning* (Davis & Chan, 2015), *test-potentiated learning* (Finn & Roediger, 2013). Pastötter and Bäuml (2014) were the first to term it *the forward effect of testing*. To keep this term concise, we termed it *the forward testing effect*.

## Appendix A: Calibration and Correlation in Experiment 3

### Calibration

To calculate List 4 calibration scores (absolute agreement between List 4 JOLs and List 4 interim test recall), the following formula was employed:

$$\text{Calibration} = \left(1 - \frac{|\text{List 4 JOL} - \text{List 4 Interim test recall}|}{12}\right) \times 100$$

Calibration scores range from 0 to 100. 0 means completely inaccurate, and 100 means completely accurate. There was no significant difference in calibration between the groups (No Interim Test group: $M = 87.91$, $SD = 12.53$; Interim Test group: $M = 84.17$, $SD = 13.22$), mean difference = 3.75, [-4.29, 11.99].

### Correlation

At the list level, for each participant in the Interim Test group, we calculated a Pearson correlation between JOLs and interim test recall across lists. The value for one participant could not be computed because of constant JOLs across lists. Average correlations were calculated via *z*-transformed scores (Silver & Dunlap, 1987). This method was also used in Experiment 4. There was no significant correlation in the Interim Test group, $r = .25$, $p = .12$.

At the participant level, for each group, we calculated a Pearson correlation between List 4 JOLs and interim test recall for that list. For the No Interim Test group, there was no significant correlation, $r = -.10$, $p = .68$, but for the Interim Test group, there was a significantly positive correlation, $r = .52$, $p = .02$. The difference in correlations between the groups was significant , $z = -1.97$, $p = .05$, revealing that the correlation between List 4 JOLs and interim test recall in the Interim Test group was greater than that in the No Interim Test group. Collapsed across groups to increase power, the correlation between List 4 JOLs and interim test recall was statistically significant, $r = .44$, $p < .01$.

## Appendix B: Calibration and Correlation in Experiment 4

### Calibration

To calculate List 5 calibration scores (agreement between List 5 JOLs and List 5 interim test performance), we applied a formula analogous to that used in Experiment 3. There was no significant difference in calibration between the groups (No Interim Test group: $M$ = 52.19, $SD$ = 37.82; Interim Test group: $M$ = 60.10, $SD$ = 30.02), mean difference = -7.91, [-29.77, 13.95].

### Correlation

For each participant in the Interim Test group, we calculated a Pearson correlation between JOLs and interim test recall across lists. There was a significant positive correlation between JOLs and interim test recall, $r$ = .41, $p$ = .01. Then for each group we calculated a Pearson correlation between JOLs and interim test recall on List 5 at the participant level. The correlations for both groups were statistically nonsignificant (No Interim Test group: $r$ = .08, $p$ = .74; Interim Test group: $r$ = .35, $p$ =.74), and there was no significant difference in correlations between the groups, $z$ = .83, $p$ = .41. Nonetheless, the correlation in the Interim Test group was numerically stronger than in the No Interim Test group, consistent with the pattern found in Experiment 3. By collapsing the data of JOLs and interim test recall on List 5 across the two groups, we observed a marginally significant Pearson correlation, $r$ = .30, $p$ = .06.

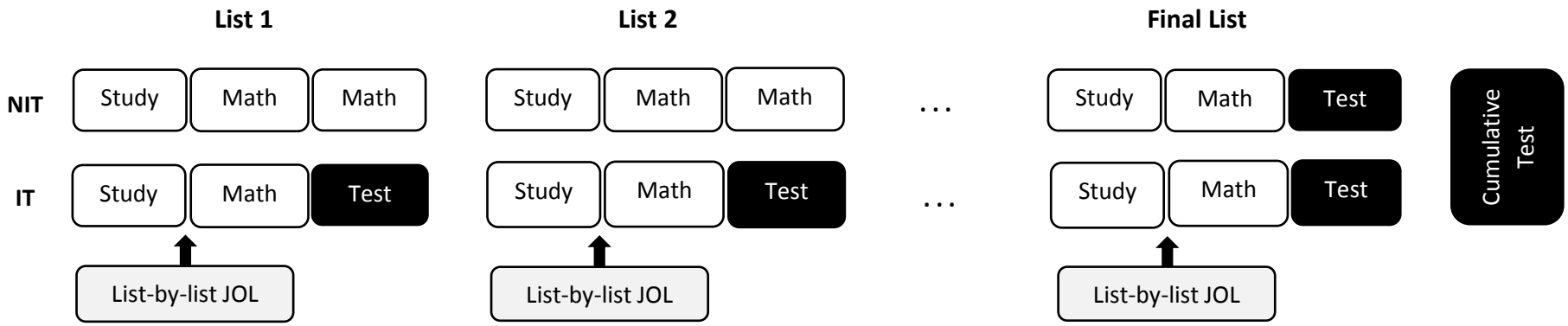


Figure 1. Experimental design schema for the No Interim Test (NIT) and Interim Test (IT) groups of Experiments 1-4. The final list was List 5 in Experiments 1 and 4 and List 4 in Experiments 2 and 3. The study materials were Euskara-English word pairs (Experiment 1), face-name pairs (Experiments 2 and 3), or word lists (Experiment 4). List-by-list judgements of learning (JOLs) were only made in Experiments 3 and 4.

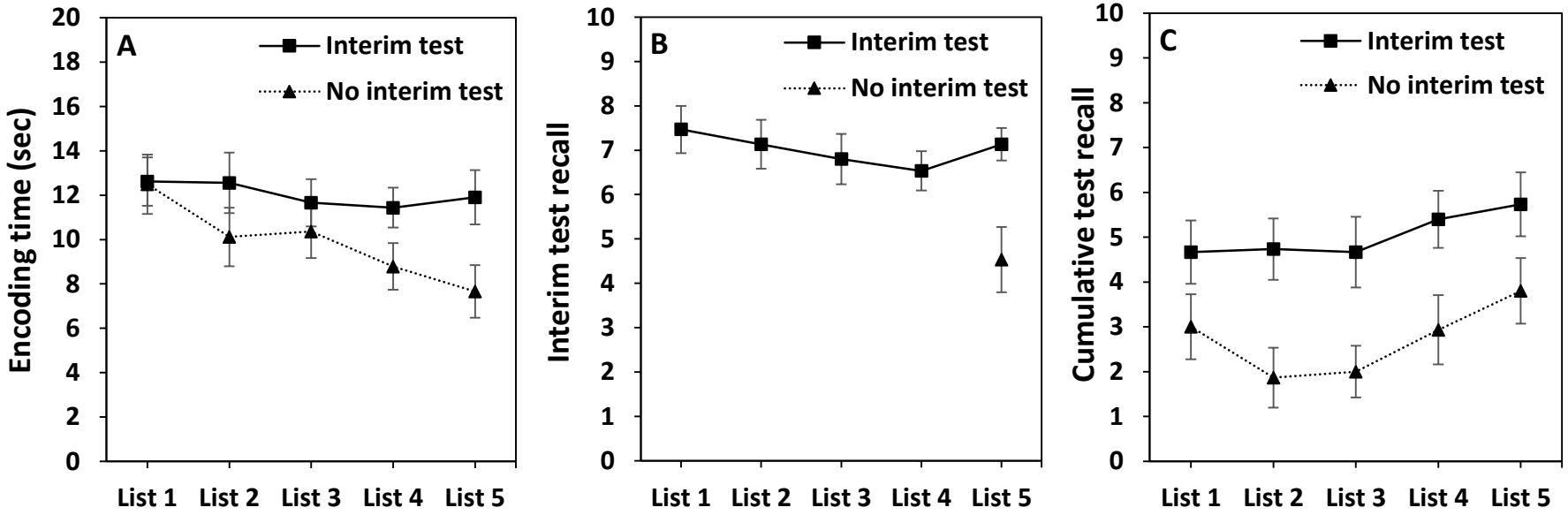


Figure 2. Experiment 1. Panel A: Time spent on encoding each Euskara-English word pair across lists. Panel B: Interim test recall across five lists. Panel C: Cumulative test recall across 5 lists. Error bars represent ± 1 standard error.

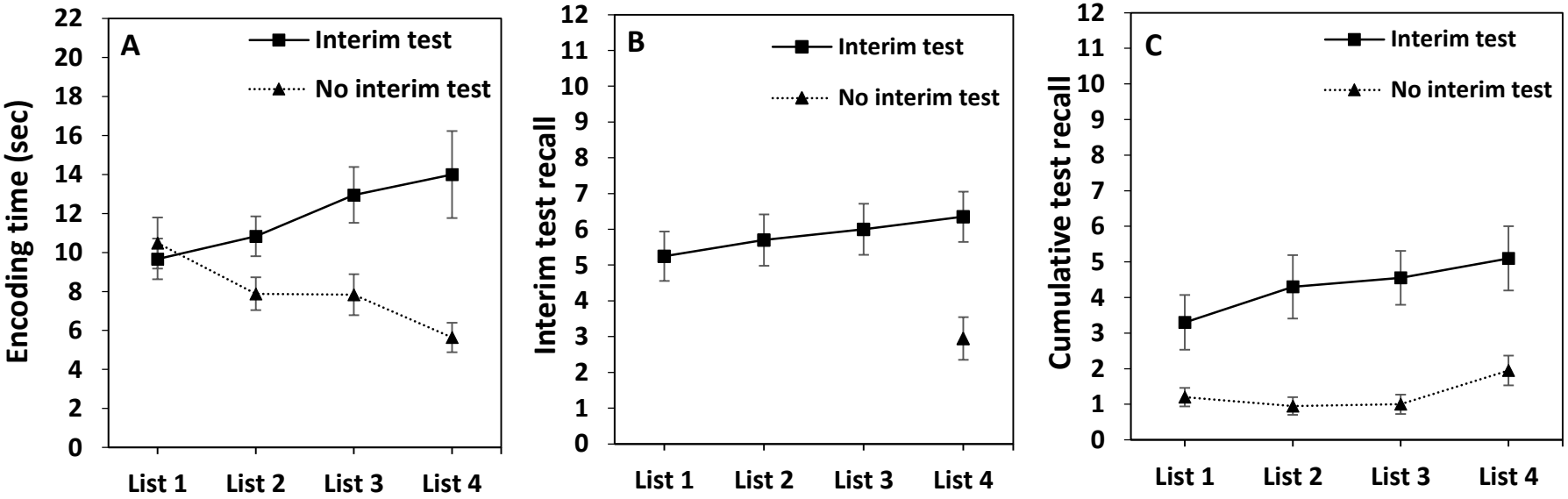


Figure 3. Experiment 2. Panel A: Time spent on encoding each face-name pair across four lists. Panel B: Interim test recall across 4 lists. Panel C: Cumulative test recall across 4 lists. Error bars represent ± 1.standard error.

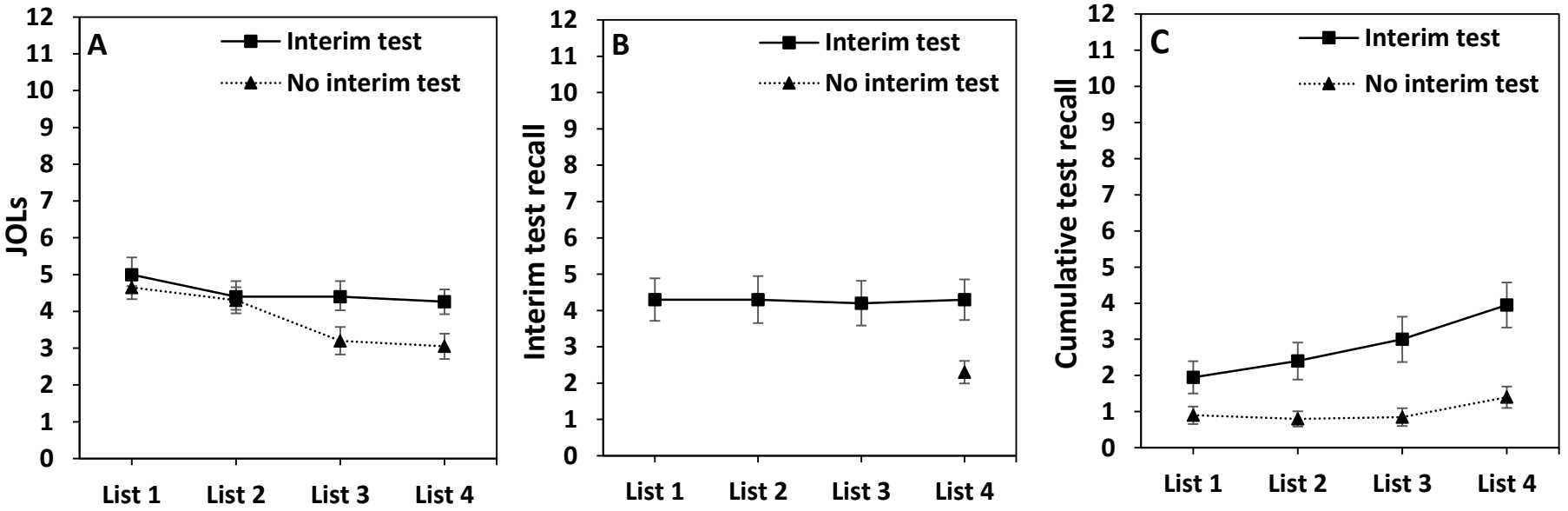


Figure 4. Experiment 3. Panel A: JOLs across four face-name lists. Panel B: Interim test recall across 4 lists. Panel C: Cumulative test recall across four lists. Error bars represent ± 1 standard error.

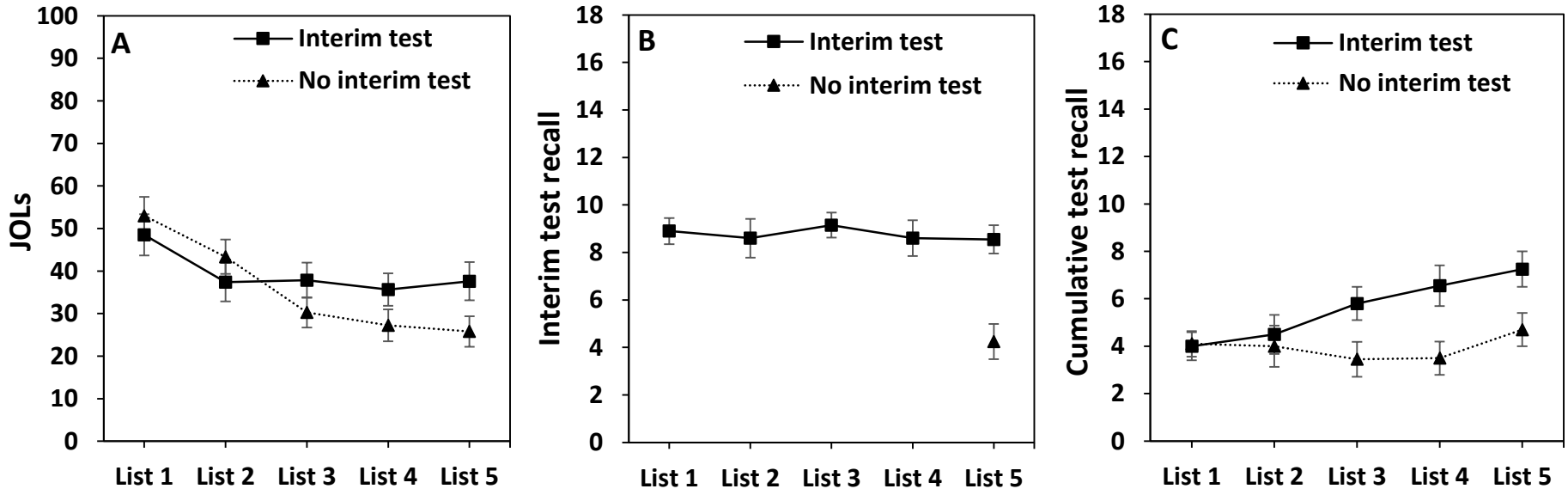


Figure 5. Experiment 4. Panel A: JOLs across five lists of words. Panel B: Interim test recall across five lists. Panel C: Cumulative test recall across five lists. Error bars represent ± 1 standard error.