Enhancing Learning and Retrieval: The Forward Testing Effect

Chunliang Yang

University College London

Thesis submitted for the degree of Doctor of Philosophy

August 2018

DECLARATION

I, Chunliang Yang, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been described in the thesis.

For the sake of reproducibility and openness, data from all experiments have been made publically available via the Open Science Framework (OSF): the data from Experiments 1-4 are available at https://osf.io/rqm6g/; from Experiments 5 and 6 at https://osf.io/3ga2t/; from Experiments 7-9 at https://osf.io/px274/; and from Experiments 10-12 at https://osf.io/24qty/. Where appropriate, experimental materials have also been uploaded to OSF. Materials derived from other sources have been indicated in the thesis.

Chunliang Yang

London, August 2018

PUBLISHED AND SUBMITTED ARTICLES

Several parts in Chapters 1 and 7 appear in Yang, C., Potts, R., & Shanks, D. R. (2018). Enhancing learning and retrieval of new information: A review of the forward testing effect. *npj Science of Learning*, *3*, 8.

Experiments 1-4 appear in Yang, C., Potts, R., & Shanks, D. R. (2017). The forward testing effect on self-regulated learning and metamemory monitoring. *Journal of Experimental Psychology: Applied*, *23*(3), 273-277.

Experiments 5 and 6 appear in Yang, C., & Shanks, D. R. (2018). The forward testing effect: Interim testing enhances inductive learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *44*(3), 485-492.

Experiments 7-9 appear in Yang, C., Chew, S., Sun. B., & Shanks, D. R. (under revision). The forward effects of testing transfer to different domains of learning. *Journal of Educational Psychology*.

My thanks for the help with data collection go to Siew-Jong Chew (Experiment 7), Bukuan Sun (Experiment 8), and Tangsheng Wang (Experiments 11 and 12).

Other research conducted during my Ph.D. but not reported in this thesis includes:

Yang, C., Hu, X., Huang, T., & Shanks, D. R. (under review). The contributions of processing fluency and beliefs to the formation of judgments of learning: A critical review. *Psychonomic Bulletin & Review*.

Yang, C., Huang, T., & Shanks, D. R. (2018). Perceptual fluency affects judgments of learning: The font size effect. *Journal of Memory and Language, 99*, 99-110.

Yang, C., Potts, R., & Shanks, D. R. (2017). Metacognitive unawareness of the errorful generation benefit and its effect on self-regulated learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *43*(7), 1073-1092.

Yang, C., Sun, B., & Shanks, D. R. (2018). The anchoring effect in metamemory monitoring. *Memory & Cognition*, *46*(3), 384-397.

ACKNOWLEDGMENTS

First and foremost my gratitude goes to my principal supervisor, Prof. David R. Shanks, for his unwavering intellectual support. It has been a great honor and fortune for me to conduct my doctoral works in his Learning, Memory, and Decision-Making Lab. I appreciate his availability at any time for my questions and concerns, the freedom to pursue my research interests, and the sharing of his scientific expertise and insights. I have learned a great deal under his intellectual supervision and it was a stimulating and pleasant experience working with him for the past a few years.

My sincere gratitude goes to my second supervisor, Rosalind Potts, for her enormous contributions to my Ph.D. research, including suggesting ways to improve my writing, constructive suggestions on experimental design and data analysis, and valuable career advice. I have deeply enjoyed her guidance and cherish the stimulating experience of working with her. My gratitude also goes to my colleagues, Tom E. Hardwick, Tina S-T. Huang, Simone Malejka, Maarten Speekenbrink, Adam Harris, Eric Schulz, Hannah Tickle, Sarah Jenkins, and Sabine Topf for their sharing of brilliant ideas, for their helpful suggestions, and for creating an enjoyable atmosphere in the lab.

Considerable thanks go to the China Scholarship Council for providing me a full scholarship to support my study. The UCL School of Life and Medical Sciences and the Experimental Psychology Society (UK) have provided financial support for my attendance at several academic conferences.

My last but greatest gratitude goes to my mom, dad, and brother. Without their selfless love and tremendous support, I could not have completed my Ph.D. study. My family has given me so much support to allow me to study somewhere so far away. I know you have sacrificed a lot and words cannot express my gratitude.

THESIS ABSTRACT

It is well established that testing of studied information, by comparison with restudying or doing nothing, enhances long-term retention of studied information – the *backward testing effect*. An accumulating body of more recent research has shown that interim testing of studied information has another important consequence: it enhances learning and retrieval of new information – *the forward testing effect*. This thesis aims to further explore the forward beneficial effects of interim testing. The research described here employs the most-widely used procedure – a multi-list method – to investigate the forward testing effect on self-regulated study time allocation (Experiments 1 and 2), metamemory monitoring (Experiments 3 and 4), inductive learning (Experiments 5 and 6), and transfer effect (Experiments 7-9). Finally, it explores whether interim tests can be used as a remedial technique to mitigate older adults' learning and memory deficits (Experiments 10-12).

Experiments 1 and 2 reveal that, in the absence of interim tests, learners systematically decrease their study times across a study phase; however, this decreasing trend is prevented (or attenuated) by interim tests. These two experiments also show that the forward benefits of interim tests generalize to self-paced learning situations. Experiments 3 and 4 show that people tend to be aware of the forward benefits of interim tests. Experiments 5 and 6 demonstrate that frequent interim tests facilitate the learning of abstract concepts, indicating that interim testing enhances inductive learning. Experiments 7-9 explore the transferability of the forward effect, in which material types (and test formats) were varied across blocks. The results confirm that the effect transfers broadly. Experiments 10-12 reveal that interim tests significantly improve older adults' learning and memory of new information. Overall, the findings shed light on the mechanisms of the forward testing effect and provide strong encouragement for learners and instructors to administer interim tests in educational contexts.

EFFECTIVE IMPACT STATEMENT

The thesis focuses on a recently developed technique for improving learning and retrieval of new information – administering interim tests during learning, which induces a beneficial *forward testing effect*.

Experiments 1 and 2 observe that administering interim tests is an effective strategy to sustain learning effort across a study phase in self-paced learning situations, and enhanced learning effort, in turn, facilitates learning and retrieval of new information. Experiments 3 and 4 document people's metacognitive insight into the forward benefits of interim testing, which may encourage learners to self-administer interim tests while learning. These two experiments also show that the forward benefits of interim tests generalize to instructor-paced learning situations. Experiments 5 and 6 demonstrate that frequent interim tests facilitate inductive learning. Experiments 7-9 establish that the forward testing effect is transferable even when material types and test formats are switched. Experiments 10-12 reveal that interim testing is an effective strategy to mitigate older adults' learning and memory deficits. Overall, the key findings provide strong encouragement for learners and instructors to administer interim tests in educational contexts.

In addition to the practical implications for improving educational practice, the findings shed significant light on the underlying mechanisms of the forward testing effect. Experiments 1 and 2 imply that both variations in the learning and retrieval processes contribute importantly to the effect. Experiments 7-9 document that frequent interim tests increase test expectancy, and a mini meta-analysis shows a positive correlation between test expectancy and recall performance. These findings support a test-expectancy explanation to account for the forward testing effect. Furthermore, Experiment 9 establishes that prior

interim tests motivate learners to exert more effort toward retrieving target information, supporting a retrieval-effort explanation.

Overall, the findings imply that interim testing is an effective strategy to sustain learning efficiency across a study phase and enhance learning and retrieval of new information.

CONTENTS

LISTS OF FIGURES

LISTS OF TABLES

CHAPTER ONE: THE BENEFICIAL EFFECTS OF TESTING

Mastering a large body of knowledge or set of skills is a considerable challenge given the limits of our cognitive resources. Ever since the founding of experimental psychology, researchers have sought to identify effective techniques to optimise learning and memory, such as providing informative feedback (Balzer, Doherty, & O'Connor, 1989), structuring materials in a spaced way (Kornell, Castel, Eich, & Bjork, 2010), administering quizzes or tests on learned information (Roediger & Karpicke, 2006b), creating concept maps (Vanides, Yin, Tomita, & Ruiz-Primo, 2005), taking notes while learning (Bohay, Blakely, Tamplin, & Radvansky, 2011), and so on (for a review, see Dunlosky, Rawson, Marsh, Nathan, & Willingham, 2013). The current thesis focuses on exploring a new technique recently developed for improving the learning and retrieval of new information – administering interim tests during learning, which induces a beneficial *forward testing effect*.

The classic (backward) testing effect

In educational settings, testing is usually regarded as an evaluative instrument to assess learning and comprehension. However, a large body of research has supplied convincing evidence that testing is also an effective technique to facilitate long-term retention of studied information (Roediger & Karpicke, 2006a). The common finding that retrieval of studied information, by comparison with restudying or doing nothing, enhances its retention, is usually termed the *testing effect* (for reviews, see Roediger & Karpicke, 2006a; Roediger, Putnam, & Smith, 2011). For reasons that will become clear, this thesis will use the term *backward testing effect* for this phenomenon. The backward testing effect has been explored in numerous experiments over the past 100 years (Abbott, 1990; Roediger & Karpicke, 2006b).

It has been well documented that the backward testing effect is a robust phenomenon across different educational materials (such as foreign-translation word pairs, text passages, lecture videos) in both the laboratory and real classrooms. Three examples are illustrative. Pyc and Rawson (2010) asked participants to study 48 Swahili-English word pairs. In a Test group, participants studied all pairs and then took three cued-recall tests. In these cued-recall tests, the Swahili words were presented one-by-one, participants were required to recall the corresponding translations, and corrective feedback was provided immediately following each recall response. By contrast, in a Restudy group, participants restudied all pairs three times following initial studying instead of taking the cued-recall tests. One week later, participants in both groups took a final cued-recall test, in which the Test group correctly recalled about three times as many translations as the Restudy group. Roediger and Karpicke (2006b) asked participants to study two text passages, with one passage studied twice and the other studied once and tested once. In a test one week later, the tested passage was substantially better recalled than the restudied one. Finally, Leeming (2002) documented the backward benefits of testing in the classroom. In a 5-week Learning and Memory course, an exam-a-day group took a short quiz following each class, in which they answered two short-answer questions, and Leeming then spent 2-3 min providing corrective feedback. The exam-a-day group took about 20 exams in total across the whole semester. By contrast, a three-exam group only took exams on three classes across the semester. In a final exam near the end of the semester, the exam-a-day group scored significantly higher than the three-exam group.

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 enhances retention of the studied information (Roediger, Putnam, et al., 2011). Besides this direct benefit, testing

also bears a variety of indirect advantages. For example, learners may interpret test results as providing feedback to diagnose the gap between their actual and desired level of learning, and then regulate subsequent study activities to narrow the perceived gap (Pyc & Rawson, 2010; Pyc & Rawson, 2012).

Another striking indirect benefit is that interim tests can facilitate 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) instructed two groups (a Test-Restudy group and a Restudy group) of participants to study 48 Swahili-English word pairs. Both groups studied the word pairs in the initial study phase and were encouraged to employ a keyword encoding strategy to remember them: generating and reporting a keyword to associate each Swahili word with its corresponding translation. Following initial study, the Test-Restudy group took three cued-recall tests, in which the Swahili words were presented one-by-one and participants recalled their translations. Immediately following each response, corrective feedback (i.e., the entire word pair) was provided for restudy and participants were allowed to change their keywords if they wished. By contrast, the Restudy group restudied all pairs three times and were allowed to change their keywords during each restudy opportunity. Pyc and Rawson observed that higher proportions of keyword shifts took place in the Test-Restudy group than in the Restudy group and that higher proportions of keywords were modified following retrieval failure versus retrieval success. Pyc and Rawson proposed that 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 efficiency and render the tested material more retrievable in future.

The above-discussed studies focused on how retrieval practice enhances memorization of specific content. Retrieval practice has been met with criticisms that it is a

"drill and kill" strategy and only enhances "inert knowledge" which cannot be utilized to solve new problems in unfamiliar contexts (Fey, 2012). However, recent research has demonstrated that retrieval of studied information not only benefits memorization of specific content but also improves knowledge transfer. For example, Butler (2010) showed that testing of a studied passage concerning bats, relative to restudying, not only enhanced retention of facts and concepts contained in that passage, but also helped students to answer inferential questions in a different knowledge domain (e.g., *What are the implications or inspirations from bat wings for designing a new type of military aircraft?*).

Besides the above-discussed backward benefits, testing of studied information also facilitates information integration, induces better knowledge organization (Zaromb & Roediger, 2010), and enhances retention of untested (but related) as well as tested information (Chan, McDermott, & Roediger, 2006). In summary, testing on studied information yields various kinds of backward benefits. (In Chapter 7 some negative effects of testing are briefly reviewed.)

Although the backward benefits of testing are broad, researchers frequently express dismay that learners and instructors tend not to appreciate these benefits, and that retrieval practice has not been applied to enhance educational practices as widely as it deserves to be (Roediger & Karpicke, 2006b). There is a range of evidence from both laboratory experiments and field questionnaires showing that instructors and learners tend to be unaware of the beneficial backward effects. For example, Roediger and Karpicke (2006b) observed that although tested passages were better recalled than restudied ones in a test one week later, participants judged that the restudied passages would be better remembered than the tested ones. In a survey conducted by Karpicke, Butler, and Roediger (2009), only 1% of participants (students from University of Washington at St. Louis) regarded retrieval practice as their best study strategy, and only 11% reported that they administered self-tests while

studying. Kornell and Bjork (2007) reported that a majority (68%) of students from the University of California, Los Angeles employed tests to determine how well they had mastered studied information, but only a minority (18%) recognized that testing effectively facilitates learning. Besides learners, instructors often minimize the use of quizzes and exams in the classroom as they believe it is time-consuming (Roediger & Karpicke, 2006a) and scoring is excessively demanding.

In summary, testing of studied information induces a variety of backward benefits: testing facilitates long-term retention; enables learners to diagnose the gap between their on-going and desired learning level, which helps them to regulate subsequent learning activities to narrow the perceived gap; facilitates knowledge transfer; produces superior information integration and knowledge organization; and enhances retention of untested (but related) information. However, both learners and instructors tend to be unaware of these backward benefits.

The forward testing effect

Many classic studies documented that learning and testing of information can accelerate the acquisition rate of new information (e.g., Schwenn & Postman, 1965; Thune, 1950). For example, Thune (1950) asked participants to study two lists of paired-associates on two experimental days. On the first day, participants studied a list and then took a test on those pairs. They then restudied the list and took another test on those pairs. This study-test cycle repeated until the recall was perfect. On the second experimental day, participants performed the same task as on the first day, except they studied a new list. Thune observed that participants required fewer cycles to reach the criterion on the second than on the first day. This facilitation effect was attributed to two psychological factors: *learning to learn* (i.e., prior learning and testing experiences teach people how to learn new information) and

warm-up (i.e., prior learning and testing experiences warm people up and prepare them to master new information).

Extending this early work, an accumulating body of recent research has established that testing of studied information, by comparison with restudying or doing nothing, can enhance learning and retrieval of new information (Pastötter & Bäuml, 2014; Pastötter, Schicker, Niedernhuber, & Bäuml, 2011; Szpunar, Jing, & Schacter, 2014; Szpunar, Khan, & Schacter, 2013; Szpunar, McDermott, & Roediger, 2008; Weinstein, Gilmore, Szpunar, & McDermott, 2014; Weinstein, McDermott, & Szpunar, 2011). Several terms have been used to refer to the fact that testing of studied information can enhance learning and retrieval of new information, such as the *interim test effect* (Wissman, Rawson, & Pyc, 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), and the *forward effect of testing* (Pastötter & Bäuml, 2014). For the sake of brevity, this thesis terms it the *forward testing effect*.

The most widely-used experimental procedure for investigating the forward testing effect is illustrated schematically in Figure 1.1. Two or three groups of participants are instructed to study a few blocks of computer-presented materials, with the materials changing from block to block. Prior to study, all participants are warned that they will take a final, cumulative test on all to-be-studied materials. They are also informed that following the study of each block the computer will decide at random whether to give them an interim test on the material contained in that block. However, in fact, the interim test decisions are predetermined. In an experimental group, interim tests are administered after every block. Here this is termed the Interim Test group. In one or two control groups, participants either take a distractor task (such as solving math problems) or restudy the material after each block

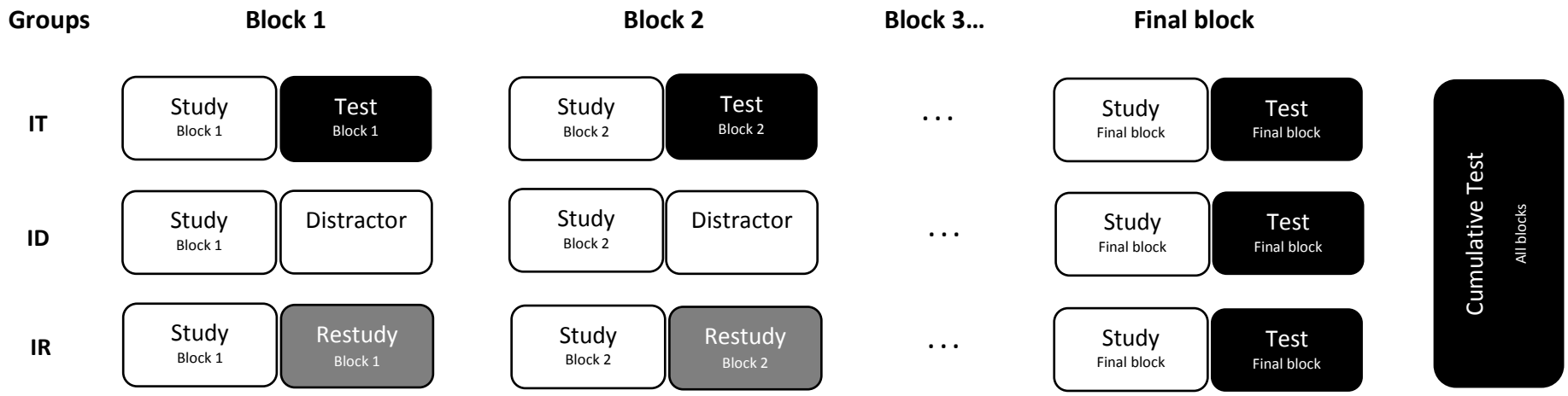

Figure 1.1: Experimental procedure for investigating the forward testing effect. The material in each block is different. The Interim Test (IT) group takes an interim test after studying each block. The Interim Distractor (ID) group completes a distractor task (e.g., solving math problems) after studying each block (except the final one) and takes an interim test on the final block. The Interim Restudy (IR) group restudies each just-studied block except the final one and takes an interim test on the final block. All groups take a final, cumulative test following the interim test on the final block.

except for the final one, on which they are tested. This thesis terms these two control groups the Interim Distractor and Interim Restudy groups, respectively.[1] Following the interim test on the final block, all three groups take a cumulative test on the material from all blocks.

The key finding is that the Interim Test group performs significantly better in the final block interim test than the control group(s). Almost all previous forward testing effect studies observed the same pattern in the cumulative test: the Interim Test group substantially outperformed the control group(s). Despite all groups studying the same material for the first time in the final block and taking an identical test on that material, learning and retrieval of that material are boosted if participants have previously been tested in preceding blocks.

Although this forward effect has been identified fairly recently, many studies have accumulated exploring its robustness and limits. It has been established that the effect is a robust phenomenon across a variety of educational materials, such as word lists (Aslan & Bäuml, 2015; Bäuml & Kliegl, 2013; Nunes & Weinstein, 2012; Pastötter et al., 2011; Pierce, Gallo, & McCain, in press; Weinstein et al., 2014), line drawings of common objects (Pastötter, Weber, & Bäuml, 2013), foreign-translation word pairs (Cho, Neely, Crocco, & Vitrano, 2016), face-name pairs (Weinstein et al., 2011), text passages (Healy, Jones, Lalchandani, & Tack, in press; Wissman et al., 2011; Zhou, Yang, Cheng, Ma, & Zhao, 2015), and lecture videos (Jing, Szpunar, & Schacter, 2016; Szpunar et al., 2013; Yue, Soderstrom, & Bjork, 2015). In summary, previous studies demonstrate that the forward testing effect reliably generalizes to many kinds of educational materials.

The forward testing effect on single item learning

[1] Several forward testing effect studies (e.g., Wissman et al., 2011) employed a different control group, in which a few blocks of materials were studied consecutively (without any interim tasks administered between blocks) prior to testing on the final block. This thesis terms this the No Interim Task group.

Szpunar et al. (2008) conducted what is by now a classic study demonstrating the forward testing effect on single item learning. In their Experiment 3, they instructed three groups of participants to study five lists of words. An Interim Test group undertook an interim test at the end of every list, in which participants were asked to freely recall the words from the just-studied list. An Interim Restudy group restudied the previous list after studying each of Lists 1-4 and took an interim test on List 5. An Interim Distractor group solved math problems after studying each of Lists 1-4 and took an interim test on List 5. Szpunar and colleagues found that the Interim Test group correctly recalled about twice as many List 5 words as the Interim Distractor and Interim Restudy groups in the List 5 interim test, which did not differ in their levels of recall. Meanwhile, the Interim Distractor and Interim Restudy groups committed about ten times as many intrusions from prior lists (proactive interference, PI; i.e., mistakenly recalling words from Lists 1-4) as the Interim Test group in the List 5 interim test. There was no difference in PI between the Interim Distractor and Interim Restudy groups.

This study clearly reveals that interim testing of studied single items enhances learning and retrieval of new items compared to no interim testing (distractor task) or restudying. This finding has been repeatedly demonstrated in recent studies using word (Aslan & Bäuml, 2015; Bäuml & Kliegl, 2013; Lehman, Smith, & Karpicke, 2014; Nunes & Weinstein, 2012; Pastötter et al., 2011; Pierce et al., in press; Weinstein et al., 2014) and picture (Pastötter et al., 2013) lists.

The forward testing effect on paired-associate learning

Following Szpunar et al. (2008), researchers began to explore the forward testing effect on paired-associate learning. For example, Weinstein et al. (2011) asked two groups of participants to study four lists of face-name pairs. An Interim Test group took an interim test

after studying each list whereas an Interim Distractor group only took an interim test on List 4. 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. The critical finding was that the Interim Test group correctly recalled about twice many names as the Interim Distractor group in the List 4 interim test. Weinstein et al. also found that the Interim Distractor group experienced substantially more PI (i.e., mistakenly recalling names from Lists 1-3) than the Interim Test group. Moreover, the forward testing effect on paired-associate learning has also been established using foreign-translation word pairs such as Swahili-English pairs (Cho et al., 2016). These studies jointly demonstrate forward testing effects on the learning of paired associates.

The forward testing effect on learning of complex materials

Researchers have also explored the forward testing effect on the learning of complex materials such as lecture videos and text passages. For example, Szpunar et al. (2013) instructed three (Interim Test/Interim Distractor/Interim Restudy) groups of participants to study an introductory statistics video, which was divided into four segments, each lasting approximately 5 min. Participants were allowed to take notes while watching the video, and they were asked to report whether their mind was "on task" (mind-wandering check) while watching the video. Szpunar and colleagues again obtained a forward testing effect in the Segment 4 interim test: The Interim Test group significantly outperformed the other two groups. They also found that the Interim Test group wrote down more notes and reported less mind-wandering than the other two groups.

Jing et al. (2016) conceptually replicated Szpunar et al.'s (2013) findings by employing a sociology lecture video as experimental material. Going beyond Szpunar et al. (2013), Jing and colleagues found that the Interim Test group reported more on-task mind-

wandering (e.g., thoughts relating the lecture content to their own experiences) and less off-task mind-wandering (zoning out). On-task mind-wandering was positively related to later memory performance whereas off-task mind-wandering was inversely related to later recall.

The forward testing effect on the learning of lecture video content has also been reported by Szpunar et al. (2014) and Yue et al. (2015) using different materials. For example, Yue et al. explored whether interim testing of a studied lecture video potentiates subsequent learning and retrieval of a new video. In their Experiment 2, Yue et al. asked two (Interim Test/Interim Restudy) groups of participants to study two scientific videos, with Video 1 concerning the life cycle of a star and Video 2 concerning lightning formation. In the Video 2 interim test, again, the Interim Test group outperformed the Interim Restudy group.

Wissman et al. (2011) explored the forward testing effect on the learning of prose passages. In their Experiment 1, they instructed participants to study a passage concerning the U.S. labor market, which was separated into three sections. An Interim Test group was tested after studying each section, whereas a No Interim Task group studied all three sections consecutively and was tested only on Section 3. In the interim tests, participants were asked to freely recall as much information as they could from the just-studied section. In the Section 3 interim test, the Interim Test group recalled about twice as much Section 3 information as the No Interim Task group. The forward testing effect on the learning of text passages has also been reported by Healy et al. (in press) and Zhou et al. (2015) using different text passages and test formats (e.g., multiple-choice tests).

Interim testing not only enhances memorization of specific content but also boosts information integration and comprehension of complex materials. For example, Jing et al. (2016) found that interim testing facilitates the integration of related information within each segment and across different segments of a lecture video. Zhou et al. (2015) explored the

forward testing effect on text comprehension. In the comprehension test, participants were required to combine a few pieces of information to answer a given comprehension question. Zhou et al. observed that the Interim Test group substantially outperformed the Interim Restudy group in the comprehension test, indicating that interim testing optimizes text comprehension.

Individual differences

In the studies reviewed above, participants were mostly college students. Can we generalize our conclusions about the forward benefits of testing to other groups? There is some evidence that the effect occurs in a range of participant groups. Pastötter et al. (2013) explored whether the forward testing effect generalizes to individuals who have suffered traumatic brain injury (TBI). TBI is associated with many memory deficits. For example, it affects short-term much more than long-term memory (Brooks, 1975). For individuals with TBI, their memory of past events (e.g., their childhood memory) is relatively intact but they suffer deficits in remembering recent events. Pastötter et al. (2013) asked TBI and healthy individuals to study three lists of line drawings of common objects. In the List 3 interim test, Pastötter et al. (2013) obtained a forward testing effect in both TBI and healthy individuals and no difference in the magnitude of the effect between healthy and TBI groups, indicating that interim testing during learning can be used to reduce memory deficits in people with TBI.

Aslan and Bäuml (2015) explored the forward testing effect in children. They asked adults, older children (average age = 8.8 y), and younger children (average age = 6.7 y) to study four lists of words. Aslan and Bäuml obtained a forward testing effect in adults and older children but not in younger children. They observed that, for older children and adults, the Interim Test groups suffered from less PI than the Interim Restudy groups in the List 4

interim test, whereas for the younger children there was no difference in PI between the Interim Test and Interim Restudy groups. Aslan and Bäuml speculated that the absence of the forward testing effect in younger children's single item learning may result from their deficits in inhibition of PI because for younger children the interim tests did not reduce the impairment from PI.

Mechanisms underlying the forward testing effect

Mechanisms that operate during either the encoding or retrieval phase (or both) may contribute to this facilitatory forward effect of interim testing and many possible explanations have been proposed to account for this effect. Here a brief review of these explanations is provided. It is important to emphasize at the outset that these accounts are not mutually exclusive, that most are at a preliminary stage of development, and that few have been subjected to direct testing of their key predictions. To aid understanding, the accounts are classified along two major dimensions, whether they regard encoding or retrieval as the main locus of the forward testing effect, and whether or not they propose that the effect is mediated by changes in motivation (see Table 1.1).

It is well-known that testing during learning can induce context changes (that is, testing modifies the mental contexts associated with studied items; see below for illustrations of the mental context changes induced by testing). Szpunar et al. (2008) postulated that the forward testing effect is mainly caused by the fact that context changes, induced by interim tests, reduce the build-up of PI and improve the recall of new information – this will be termed the *release from PI* theory. Interim testing of studied items updates these items' mental contexts, and hence these studied/tested items are associated with both a study (S) and a retrieval (R) context (Karpicke, Lehman, & Aue, 2014). Following the interim tests on studied items, participants study some new items, which are only associated with a study

Table 1.1: Theories proposed to account for the forward testing effect

Theories	References	Descriptions	Motivational?	Active phases
Release from PI	Szpunar et al. (2008)	Interim testing induces context changes between blocks, which reduce the build-up of PI and facilitate recall of target (new) items.	No	Retrieval
Encoding reset	Pastötter et al. (2011)	Interim testing induces context changes between blocks, which "reset" subsequent encoding of new information and make it as effective as the encoding of prior information.	No	Encoding
Activation facilitation	Wissman et al. (2011)	Interim testing induces greater retention of tested information and makes the tested information more active while encoding new information, which helps encoding and comprehension of new information.	No	Encoding
Encoding strategy	Cho et al. (2016)	Interim testing induces more effective encoding strategies than no interim testing.	No	Encoding
Retrieval strategy	Cho et al. (2016)	More effective retrieval strategies are developed during prior interim tests, which facilitate recall of target (new) items in the subsequent interim test.	No	Retrieval

Test expectancy	Weinstein et al. (2014)	Interim testing induces a greater expectancy of an immediate interim test, which motivates more effort toward encoding new information.	Yes	Encoding
Failure-encoding-effort	Cho et al. (2016)	Retrieval failures in prior interim tests induce dissatisfaction and motivate more effort toward encoding new information.	Yes	Encoding
Retrieval effort	Cho et al. (2016)	Retrieval failures in prior interim tests motivate more effort to retrieve the target (new) items in the subsequent interim test.	Yes	Retrieval

(S) context. In the subsequent interim test in which they are required to recall the target (new) items, the context difference between the previous items, which have been both studied and tested (and associated with both contexts S and R), and the new items (only associated with context S) facilitates differentiation between these items and reduces the impairment from PI. The above-reviewed studies, which explored the forward testing effects on single item learning and paired-associate learning, offer strong support for the release from PI theory: These studies observed that interim testing reduces the build-up of PI. Besides, Bäuml and Kliegl (2013) provided further evidence to support this contextual list segregation conjecture. They asked participants to study three lists of words. The results showed that, by comparison with restudying, interim tests on Lists 1 and 2 significantly enhanced recall of List 3 and reduced response latencies. Shorter response latencies imply a smaller memory search set, consistent with more effective discrimination between the target and non-target lists.

Different from the release from PI theory, which focuses on the influence of context changes on subsequent recall of new information, an *encoding reset* theory, proposed by Pastötter et al. (2011), focuses on the influence of context changes on the subsequent encoding of new information. Specifically, this theory postulates that interim tests induce context changes between blocks, which in turn induce a "reset" of subsequent encoding, making it as effective as the encoding of material in prior blocks. Indeed, Pastötter, Bäuml, and Hanslmayr (2008) found that an imagination task (e.g., participants imagined walking through their parents' living room), which induces mental context changes between the studying of two lists of words, makes the learning of the second list as effective as that of the first list. In contrast, in the absence of the imagination task, less attention is attached to the encoding of the second list compared to the encoding of the first one.

Both the release from PI and encoding reset theories focus on the roles of context changes in the forward testing effect. Nonetheless, both theories have difficulty explaining

the forward testing effect observed in an important study by Wissman et al. (2011). In their Experiment 4, Wissman and colleagues had three groups of participants study a three-section passage. An Interim Test group took a free recall test after studying each section, an Interim Distractor group solved math problems after studying each of Sections 1 and 2 but took a free recall test on Section 3, and a Section-3 group only studied Section 3 and took a free recall test on it. The results showed that the Interim Test group recalled about twice as much information from Section 3 as the Interim Distractor and Section-3 groups. This is striking because according to the release from PI theory, recall in the Section-3 group should be better or at least equal to that in the Interim Test group, for whom at least some PI would accumulate across the study phase. Similarly, according to the encoding reset theory, recall in the Section-3 group should be better or at least equal to that in the Interim Test group, because the Section-3 group's encoding effectiveness of Section 3 material should be at least as effective as that of the Interim Test group. However, recall in the Section-3 group was, in fact, poorer than in the Interim Test group.

Hence, Wissman et al. proposed an *activation facilitation* theory to account for their forward testing effects. This theory postulates that greater activation and retention of learned information, induced by prior interim tests, can facilitate encoding of new related information, especially for complex materials such as lecture videos and text passages. Different sections of a passage or a lecture video are related. Interim testing of prior sections improves retention of tested information compared to restudying or doing nothing. While encoding the target section (new information), the tested information is more activated and accessible on this theory, which in turn facilitates comprehension of new information (Bransford & Johnson, 1972).

Besides the three theories discussed above, Cho et al. (2016) proposed that the forward testing effect may be produced by encoding strategy changes – the *encoding strategy*

theory. This theory hypothesizes that prior interim tests inform people what kind of test to expect and accordingly they adjust their encoding strategies, which facilities subsequent encoding of new information. Previous studies have shown that testing can foster the development and adoption of more effective learning strategies (Pyc & Rawson, 2010; 2012; Soderstrom & Bjork, 2014). For example, Soderstrom and Bjork (2014) found that, following testing, individuals are likely to employ more effective encoding strategies (e.g., relating the item to something that is meaningful to them) than they are following restudying.

Cho et al. (2016) also proposed a complementary *retrieval strategy* theory to account for the forward testing effect, which postulates that prior interim tests on studied information help people to adopt more efficient retrieval strategies. Specifically, this theory postulates that participants gradually develop more effective retrieval strategies across successive interim tests (for an illustration that prior interim tests induce retrieval strategy changes, see Thomas & McDaniel, 2013), and these more effective retrieval strategies facilitate recall of new items in the subsequent interim test.

This section has summarised five (release from PI; encoding reset; activation-facilitation; encoding strategy; retrieval strategy) theories explaining the forward testing effect, and these theories focus on the roles of non-motivational factors in the forward testing effect. In contrast, three other theories assume that the forward testing effect is caused or mediated by the fact that prior interim tests motivate people to exert greater effort toward encoding/retrieval of new information. For example, Weinstein et al. (2014) suggested that the forward testing effect is attributable to test expectancy. Here this is termed the *test expectancy* theory. This theory postulates that, since the Interim Test group is always tested on prior blocks, they should have a high expectancy of an interim test on the next list, and high test expectancy motivates people to exert greater effort toward encoding new information (Agarwal & Roediger, 2011; Weinstein et al., 2014). To test this idea, Weinstein

et al. asked an Interim Test and an Interim Distractor group to study five lists of words. Before studying each list, all participants were instructed to report how likely they thought it was that they would be asked to take an immediate interim test on the next list. The results showed that the Interim Test group's test expectancy increased whereas the Interim Distractor group's decreased across lists. A variety of studies have established that expecting a later test improves subsequent learning effectiveness and test performance (e.g., Agarwal & Roediger, 2011; Eitel & Kühl, 2015; Middlebrooks, Murayama, & Castel, in press; Nestojko, Bui, Kornell, & Bjork, 2014).

An alternative explanation for why interim tests motivate people to devote greater encoding effort was proposed by Cho et al. (2016), who postulated that it is retrieval failures in prior interim tests that motivate people to commit more encoding effort. Here this is termed the *failure-encoding-effort* theory. Retrieval failures in prior interim tests induce dissatisfaction with prior learning as well as awareness of the difficulty of achieving successful recall, leading to enhanced study effort to mitigate the dissatisfaction. Consistent with this idea, previous studies have shown that retrieval failures or committing errors in prior tests can potentiate subsequent encoding through enhancing curiosity and learning motivation (Kornell, Hays, & Bjork, 2009; Potts, Davies, & Shanks, 2018; Potts & Shanks, 2014; Yang, Potts, & Shanks, 2017).

Finally, besides enhanced encoding effort (induced by enhanced test expectancy and/or retrieval failures in prior interim tests), enhanced retrieval effort may play a role in the forward testing effect (Cho et al., 2016). For instance, Cho et al. (2016) attributed the forward benefit of interim testing to enhanced retrieval effort – the *retrieval effort* theory: Retrieval failures in prior interim tests induce dissatisfaction about prior interim test performance and then motivate participants to exert more retrieval effort in the subsequent interim test to

alleviate their dissatisfaction. This theory has not been tested yet and the thesis will return to this in Chapter 6.

Overall, this section has discussed at least eight possible theories, each proposing a mechanism that may underlie the forward testing effect. Some theories are similar. For example, the retrieval effort theory can be regarded as a subset of the retrieval strategy theory, as committing more effort during retrieval is a form of retrieval strategy change. Similarly, committing more effort during encoding can be seen as a form of encoding strategy change. Cho et al. (2016) noted that encoding or retrieval effort change is a quantitative change whereas encoding or retrieval strategy change is a qualitative change. Quantitative changes refer to the changes in the magnitude of effort people devote to a task, whereas qualitative changes mainly refer to changes in the ways people encode and retrieve items. Therefore, the thesis discusses encoding/retrieval-effort and encoding/retrieval-strategy theories separately.

These eight theories can be divided into two clusters according to the phase at which the putative mechanism they describe is assumed to be active: encoding (encoding reset; activation facilitation; encoding strategy; test expectancy; failure-encoding-effort) and retrieval (release from PI; retrieval strategy; retrieval effort). They can also be divided according to their assumptions about the role of motivation: motivational (test expectancy; failure-encoding-effort; retrieval effort) and non-motivational (release from PI; encoding reset; activation facilitation; encoding strategy; retrieval strategy). Although this thesis divides the mechanisms these theories propose into different clusters, it is important to reiterate that they need not be mutually exclusive and some of them may operate in parallel in some situations and produce overlapping forward testing effects.

For different materials and in different situations, some mechanisms may play the main roles and others may play lesser or even no role. For example, for complex materials (e.g., passages, lecture videos), there may be no PI and therefore the release from PI mechanism may play little or no role (Wissman et al., 2011). In contrast, the activation facilitation mechanism may play an important role in the learning of complex materials (Wissman et al., 2011). For single items (e.g., unrelated word lists), activation and retention of prior information cannot aid comprehension of new information and therefore the activation facilitation mechanism may play a small or no role, while the release from PI mechanism may play a more important role (Szpunar et al., 2008). These findings imply that, in different contexts and for different materials, forward testing effects may be attributed to distinct cognitive and/or metacognitive mechanisms.

Rationale of the thesis

Given that the forward testing effect was identified fairly recently, many aspects of this important effect remain poorly understood and further investigation of the robustness, boundaries, and underlying mechanisms of this effect are needed. Without a much deeper exploration of its mechanisms and boundary conditions, educational translation and exploitation will be hindered. Hence, the experiments contained in this thesis were designed to address this.

All previous studies explored the forward testing effect in experimenter- or instructor-paced situations. But of course, the pace of studying is often self-determined. The question of whether or not interim tests can influence self-regulated learning of new information has not been explored yet. Therefore, Experiments 1 and 2 were designed to investigate whether or not interim tests can modify self-regulated study time allocation across a study phase and improve retrieval of new information in a self-paced study situation (see Chapter 2).

Prior research has documented that people tend to be unaware of the backward testing effect (Agarwal, Karpicke, Kang, Roediger, & McDermott, 2008; Kornell & Son, 2009; Roediger & Karpicke, 2006b). This alignment between an objective benefit on the one hand and metacognitive awareness on the other cannot be taken for granted. To date, no research has explored whether people are aware of the forward testing effect. Therefore, Experiments 3 and 4 were designed to explore whether people tend to be aware of the forward benefits of interim testing (see Chapter 3).

Previous studies investigating the forward testing effect were largely restricted to exploring the forward testing effect on low-level learning – learning and remembering specific content (Pastötter & Bäuml, 2014). It is unknown whether testing can have a facilitatory forward effect on high-level learning, such as inductive learning. Different from low-level learning (i.e., remembering specific items), inductive learning requires summarizing the characteristics shared by multiple exemplars and applying prior knowledge when making uncertain inferences about the environment that go beyond direct experience. Inductive learning is a key element of how humans learn and understand the world and a key component of formal education. Schacter and Szpunar (2015) suggested that "An important question is whether interpolated retrieval/testing also enhances learning at a conceptual level" (p. 67). Therefore, Experiments 5 and 6 filled this gap by exploring the forward testing effect on the learning of painting styles (see Chapter 4).

In all previous forward testing effect studies, the type of material has always been the same across lists/blocks, and whether the forward testing effect is transferable across different domains of learning has not been addressed. It is important to explore whether testing of studied information from one domain can facilitate learning and retrieval of new information from a *different* domain - the transferability of this effect - because, in natural learning situations, the types of to-be-studied materials are frequently switched. For example,

high school students may take a history class, then a geography class, and then a biology class. Even within a class, the content frequently varies (e.g., learning a concept or definition, then learning about a technique). Experiments 7-9 were designed to explore the transferability of the forward testing effect (see Chapter 5).

Learning and memory deficits are general complaints amongst older adults (Emery et al., 2008; Ikier, Yang, & Hasher, 2008; Tse, Balota, & Roediger, 2010). Participants in previous forward testing effect studies were largely restricted to college students and it has never been explored whether interim tests can be profitably used to mitigate older adults' learning and memory deficits. Therefore, Experiments 10-12 were designed to explore whether interim tests can be used as a remedial technique for older adults' learning and memory deficits (see Chapter 6).

Finally, the thesis summarizes the empirical findings on the forward testing effect observed in this thesis and other studies, and discuss practical principles for improving learning and education. The possible negative effects of administering interim tests on the learning of new information and how to mitigate such negative effects are also discussed. The final discussion also makes some suggestions for future research to further investigate aspects of this important effect that are currently poorly understood (see Chapter 7).

CHAPTER TWO: THE FORWARD TESTING EFFECT ON SELF-REGULATED STUDY TIME ALLOCATION

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 behavior to optimize learning, comprehension, and knowledge transfer. However, recent studies revealed 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 using interim tests to optimize self-regulated learning of new information. One aim of this chapter is to fill this gap. Specifically, this chapter explores how interim tests influence subsequent self-regulated study time allocation when learning new information.

Self-regulated learning and testing

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 an honoring condition, participants reviewed the pairs which were selected to be restudied. In contrast, participants in a dishonoring condition reviewed the pairs which they had not selected for restudy. In a later test, participants in the honoring condition significantly outperformed those in the dishonoring condition. This study revealed that people could 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 retention. In some situations, self-regulated learning impairs retention compared with experimenter-paced learning. For instance, Kornell and Bjork (2008b) allowed some participants to remove some Swahili-English pairs from further study when they thought they 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 (2008b) 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-phone*), there was no semantic association from the cue to the 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 a Test group undertook a cued recall test and 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 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. Experiments 1 and 2 were designed to fill this gap.

Rationale of Experiments 1 and 2

In all previous research investigating the forward testing effect, the initial encoding phase was always experimenter-paced, which is not common outside the formal classroom. The question of whether or not interim tests can influence self-regulated learning of new information is yet to be answered. 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.

Besides practical implications, exploring the forward testing effect on self-regulated study time allocation also bears theoretical implications. As summarised in Table 1.1, mechanisms that operate during either the encoding or retrieval phase (or both) may contribute to this facilitatory forward effect of interim testing. Experiments 1 and 2 investigated whether interim tests can modify study time allocation across lists and improve retention of new information. Directly measuring self-regulated study time allocation will 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 variations in retrieval processes constitute another possible source of this effect.

Experiment 1

Experiment 1 was conducted to determine how interim tests influence subsequent encoding time allocation when learning new information. In previous research, the effect has been studied only under experimenter-paced conditions. Another aim, therefore, was to determine whether the forward testing effect can be replicated when the encoding procedure

is self-paced, which is more typical of self-regulated learning. The University College London (UCL) Department of Experimental Psychology provided ethical approval for Experiments 1-10 in this thesis.

Method

Participants

According to the effect size (Cohen's d = 1.78) in a previous related study (Weinstein et al., 2011), a sample size calculation was conducted using G*power (Faul, Erdfelder, Lang, & Buchner, 2007), which showed that about eight participants in each group were required to observe a significant (α = .05) forward testing effect at 0.9 power. Given that the principal stimuli (foreign-translation pairs) in this experiment were different from the face-name pairs employed in Weinstein et al. (2011), the sample size was increased to 15 in each group. Thirty participants, 24 females, with an average age of 24.10 years (SD = 7.22) were recruited from the UCL participant pool (UCL SONA system). Their first language was English. All of them were naïve to the aim of the experiment and reported normal or corrected-to-normal vision and no prior experience of Euskara, the language of the Basque region of Spain. 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 (e.g., *sagu – mouse*) were selected from a set constructed by Potts and Shanks (2014). These 50 Euskara nouns were divided into 5 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 *Psychtoolbox* 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 experiment design schema in Figure 2.1). In this study, the No Interim Test group continued solving math problems rather than restudying the items following each of Lists 1-3. If participants in the No Interim Test group restudied the items following each list, they would come to expect restudying opportunities and this might decrease their initial encoding effort. For example, Sparrow, Liu, and Wegner (2011) tested the effect of saving information on a computer. For the information erased from the computer, participants' recall in a later test was significantly better than for information saved on the computer, presumably because participants expected that they could access the saved information later, which thus reduced the need to fully encode the saved information.

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*

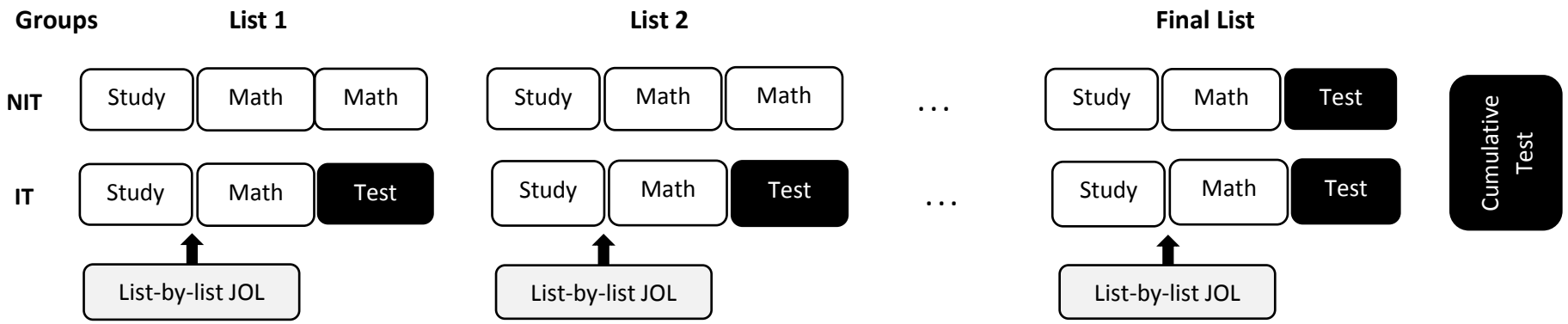

Figure 2.1: Experimental design schema for the No Interim Test (NIT) and Interim Test (IT) groups of Experiments 1-4. The final list was List 5 in Experiments 1 and 4, and List 4 in Experiments 2 and 3. The study materials were Euskara-English word pairs (Experiment 1), face-name pairs (Experiments 2 and 3), or word lists (Experiment 4). List-by-list JOLs were only made in Experiments 3 and 4.

+ 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. The experiment lasted about 35 min.

Results

Encoding time

The mean encoding time per word pair on each of Lists 1-5 for both groups is shown in Figure 2.2A. These data were analyzed 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, $F(1, 28) = 14.41$, $p < .01$, $\eta_p^2 = .34$, and 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, $F(1, 28) = 1.92$, $p = .177$. 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, $F(4, 56) = .74$, $p = .57$.

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 significantly decline across lists. An independent-samples *t* test revealed that participants in the Interim Test group spent

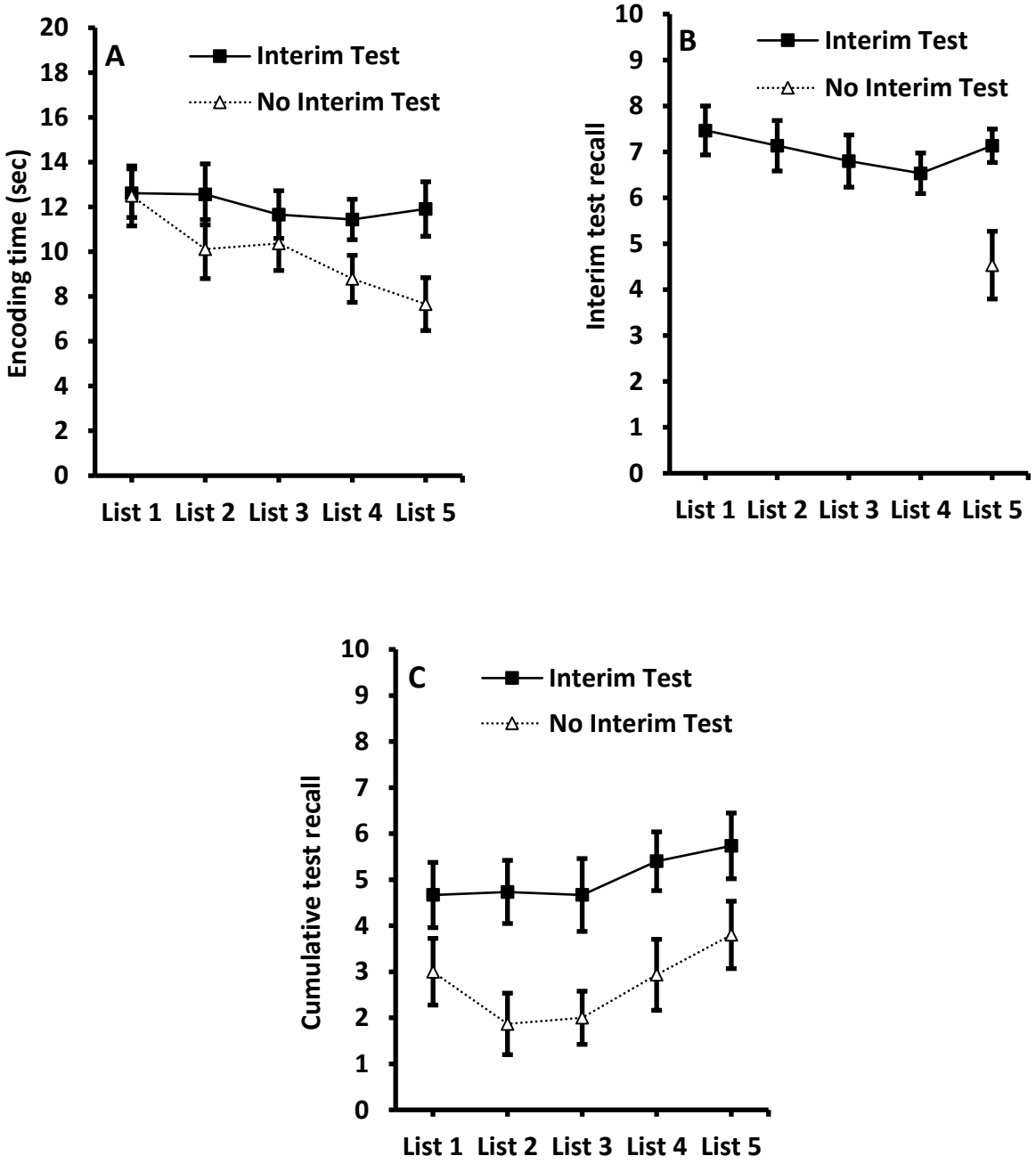

Figure 2.2: Experiment 1. Panel A: Time spent on encoding each Euskara-English word pair across five lists. Panel B: Interim test recall across five lists. Panel C: Cumulative test recall across five lists. Error bars represent ± 1 standard error.

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) = [0.66, 7.84], Cohen's d = 0.97. 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 2.2B 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 significantly fluctuate across lists, $F(4, 56) = 1.11$, $p = .36$. The critical comparison of interim test recall between the groups was on List 5. Levene's test showed that the assumption of homoscedasticity 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 translations, 95% CI = [0.83, 4.37], Cohen's $d = 1.12$.

By comparison with the No Interim Test group, the Interim Test group recalled more words correctly and hence fewer opportunities were left for intrusions (mistakenly recalling another word's translation from any list including the current list) in the List 5 interim test. However, 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 translations, 95% CI = [-0.24, 2.24], Cohen's $d = 0.60$. Critically, when the analysis was 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 ($M = 0.60$, $SD = 1.45$), mean difference = 1.07 translations, 95% CI = [.02, 2.11], Cohen's d = 0.77. No significant difference in intrusions from the current list

between the groups was detected (No Interim Test group: M = 0.80, SD = 0.77; Interim Test group: M = 0.87, SD = 1.06), mean difference = -0.07 translations, 95% CI = [-0.76, .63], Cohen's d = 0.07. 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. The data were separately analyzed 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 standard backward testing effect, as these items were tested after each list in one group but not the other. The theoretical analysis of List 5 recall also includes 2 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 2.2C, participants in the Interim Test group recalled more List 1-4 translations than participants in the No Interim Test group, mean difference = 9.81 translations, 95% CI = [2.63, 16.98], Cohen's d = 1.05. This is a very large difference, roughly a doubling of the number of targets recalled. Both of the factors mentioned above seem to be contributing. The group difference is evident on List 1, where study time is the same across groups; this is a standard (backward) testing effect. 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 was only marginally significant, 95% CI = [-.11, 4.24], Cohen's *d* = 0.69.

Correlations between study time and interim test recall

For each group, a Pearson correlation was calculated 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, $r = .45$, $p = .01$.

Summary

In the absence of interim tests, participants decreased their encoding time across lists. In contrast, encoding time did not decrease significantly 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 claim that variations in the encoding processes contribute to the forward testing effect: 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 claim that variations in retrieval processes contribute to the forward testing effect: interim 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, four lists of 12 face-name pairs were employed as the experimental materials, as did Weinstein et al. (2011). Experiment 2 explores whether or not the forward testing effect on self-regulated study time allocation extends to face-name learning.

Method

Participants

According to the effect size of the forward testing effect (Cohen's d = 1.12) in Experiment 1, sample size calculation was conducted using G*power (Faul et al., 2007), which showed that about 18 participants in each group were required to observe a significant (α = .05) forward testing effect at 0.9 power. 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 and they 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 at 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 at 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: five participants in each group studied these lists in each of four 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 four 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 2.1). 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. The experiment lasted about 30 min.

Results

Encoding time

The mean encoding time per face-name pair on each of Lists 1-4 for both groups is shown in Figure 2.3A. These data were analyzed 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, $F(3, 114) = .48$, $p = .69$, 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 sec per pair, 95% CI = [0.27, 5.62], Cohen's $d = 0.70$), 3 (mean difference = 5.11 sec per pair, 95% CI = [1.51, 8.71], Cohen's $d = 0.91$), and 4 (mean difference = 8.35 sec per pair, 95% CI = [3.58, 13.12], Cohen's $d = 1.21$) than those in the No Interim Test group. There was no significant difference in List 1 encoding time between the groups, mean difference = -0.81 sec per pair, 95% CI = [-4.20, 2.58], Cohen's $d = 0.15$.

Interim test recall and intrusions

Figure 2.3B 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 names, 95% CI = [1.74, 5.06], Cohen's $d = 1.31$, 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

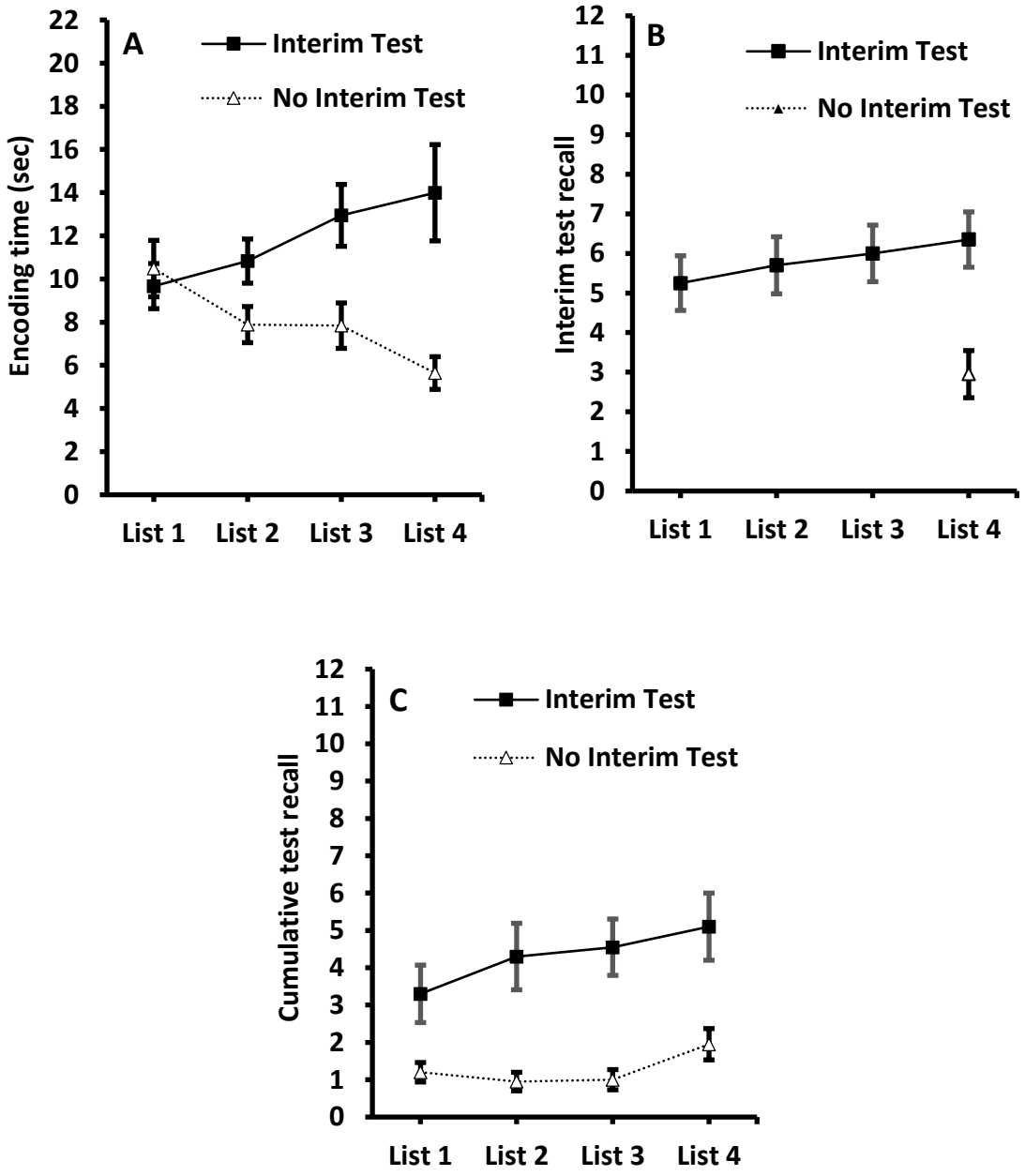

Figure 2.3: Experiment 2. Panel A: Time spent on encoding each face-name pair across four lists. Panel B: Interim test recall across four lists. Panel C: Cumulative test recall across four lists. Error bars represent ± 1 standard error.

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 names, 95% CI = [-0.21, 3.31], Cohen's d = 0.57. 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 names, 95% CI = [0.17, 2.23], Cohen's d = 0.75. The two groups made roughly equivalent numbers of current list intrusions (mistakenly recalling another face's name from the current list, No Interim Test: M = 2.80, SD = 3.00; Interim Test group: M = 2.45, SD = 2.21), mean difference = 0.35 names, 95% CI = [-1.34, 2.04], Cohen's d = 0.13. 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 2.3C, 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 names, 95% CI = [5.27, 12.73], Cohen's d = 1.54. Consistent with the forward testing effect observed in the List 4 interim test, participants in the Interim Test group recalled more List 4 names than participants in the No Interim Test group in the cumulative test, mean difference = 3.15 names, 95% CI = [1.46, 4.84], Cohen's d = 1.19.

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, $r = .30$, $p = .06$.

Summary

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 the claim that both encoding and retrieval factors play roles in the forward testing effect.

Discussion

Across two experiments, the forward testing effect was replicated when the encoding procedure was self-paced. In the final list interim test, participants in the Interim Test groups outperformed those in the No Interim Test groups. This effect was substantial and amounted to an approximate doubling of the final list interim test recall. Both experiments showed a decreasing slope of encoding time across lists in the No Interim Test groups, which was not present in the Interim Test groups (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. Both experiments showed 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.

Participants in the Interim Test groups spent more time encoding the final list than participants in the No Interim Test groups, 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). At the same time, the release from PI in the Interim Test groups observed in both experiments supports the alternative but not mutually exclusive idea that facilitation of retrieval is partly responsible for the forward testing effect (Szpunar et al., 2008). In both experiments, in the final list interim test higher proportions of intrusions were from the current list in the Interim Test groups than in the No Interim Test groups, indicating that the memory search set in the Interim Test groups was smaller and that these participants were better able to control their recall from the current list (Weinstein et al., 2011), which again supports the role of retrieval factors in the forward testing effect.

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 et al., 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 increased 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 encourage people to focus their attention and effort on encoding new information. In 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.

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 participants reported any prior study experience of Euskara. It seems reasonable therefore to speculate that face-name encoding is more fluent than Euskara-English encoding. On the basis of their experienced fluency, if 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 better appreciate the gap between their perceived and actual learning status. As Chapter 3 will show, 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 their encoding time when studying face-name pairs in Experiment 2 but not when studying 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.

In the final cumulative test, in both experiments, the Interim Test groups significantly outperformed the No Interim Test groups. The superior cumulative performance in the Interim Test group can be partially attributed to the backward testing effect (Roediger & Karpicke, 2006a, 2006b; Weinstein et al., 2011) because items were initially tested in the Interim Test group but not in the No Interim Test group (except final list items). 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 (Pastötter et al., 2011; Szpunar et al., 2013).

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. Experiments 1 and 2 established that interpolated testing maintains people's motivation to commit study time to encoding new information, which enhances learning and retrieval of new information. 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. Experiments 1 and 2 showed that interpolated testing can prevent intrusions from prior learning segments.

In 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. These

findings lead to a strong recommendation that interim tests can be profitably used to promote the learning of new information when learning is self-paced.

CHAPTER THREE: THE FORWARD TESTING EFFECT ON METAMEMORY MONITORING

Experiments 1 and 2, together with prior forward testing effect studies (Cho et al., 2016; Pastötter et al., 2011; Szpunar et al., 2013; Szpunar et al., 2008; Weinstein et al., 2014; Weinstein et al., 2011), suggest that learning and retrieval of new information can be considerably boosted by interim testing on 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 et al., 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 testing effect. In contrast, if they appreciate the forward benefits of interim testing, their motivation to self-administer interim tests may be boosted.

To date, only one study has employed the multi-list method to investigate forward and backward testing effects on metamemory calibration. Szpunar et al. (2014) divided an online statistics lecture into four 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 judgment of learning (JOL) 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, while a single test did not enhance recall, it did reduce the level of predicted recall. In this study, participants' global JOLs might have been 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 Experiments 3 and 4, participants were asked to make a JOL on each separate list to estimate their performance in a possible future interim test. By directly measuring changes in JOLs across successive lists, this chapter asks 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, this chapter is able to ask (c) whether or not JOLs are sensitive to the forward benefits of interim testing.

Experiment 3

The primary aim of Experiment 3 was to explore people's metacognitive insight regarding this forward testing effect. In the previous two experiments (Experiments 1 and 2), participants in the No Interim Test group decreased their encoding time across lists. In contrast to the previous two experiments, in Experiment 3, the encoding procedure was experimenter-paced. Experiment 3, therefore, investigated whether participants in a No Interim Test group appreciate the decrease of their learning effectiveness across lists when the encoding phase is experimenter-paced. Previous studies showed that asking participants to make JOLs affects their self-regulated study time allocation and that measuring metamemory changes memory. For example, when expecting to make JOLs people may sacrifice some of their study time to assess the memorability of an item (Mitchum, Kelley, & Fox, 2016). In self-paced conditions, it is reasonable to assume that participants would take JOLs as a basis for deciding when to terminate study and would stop their studying when

their assessment of learning (JOL) reaches an acceptable threshold level. To directly explore the forward testing effect in metamemory monitoring, Experiment 3 therefore employed an experimenter-based procedure.

Method

Participants

The same sample size was employed as in Experiment 2. 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 and they 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. After studying each individual list (and before the test on that list for the Interim Test group), 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 will not recall any names correctly*") to 12 ("*I will recall all names correctly*") (see the experiment design schema in Figure 2.1). The experiment lasted about 25 min.

Results

JOLs

Average JOLs on each of Lists 1-4 for both groups are shown in Figure 3.1A. These data were analyzed 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, $F(1, 38) = 2.85$, $p = .10$. 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, $F(3, 57) = 1.14$, $p = .34$.

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 names, 95% CI = [0.15, 2.25], Cohen's d = 0.72) and 4 (mean difference = 1.15 names, 95% CI = [0.27, 2.03], Cohen's d = 0.84). No statistically significant difference between the two groups' JOLs on Lists 1 and 2 was detected, $0.19 \leq |t|\text{s} \leq 0.45$, $.65 \leq p\text{s} \leq .85$.

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 3.1B. For the Interim Test group, the data were analyzed 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 the Huynh-Feldt correction was applied. The ANOVA revealed no main effect of List, $F(2.13, 40.38) = .02$, $p = .98$, 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

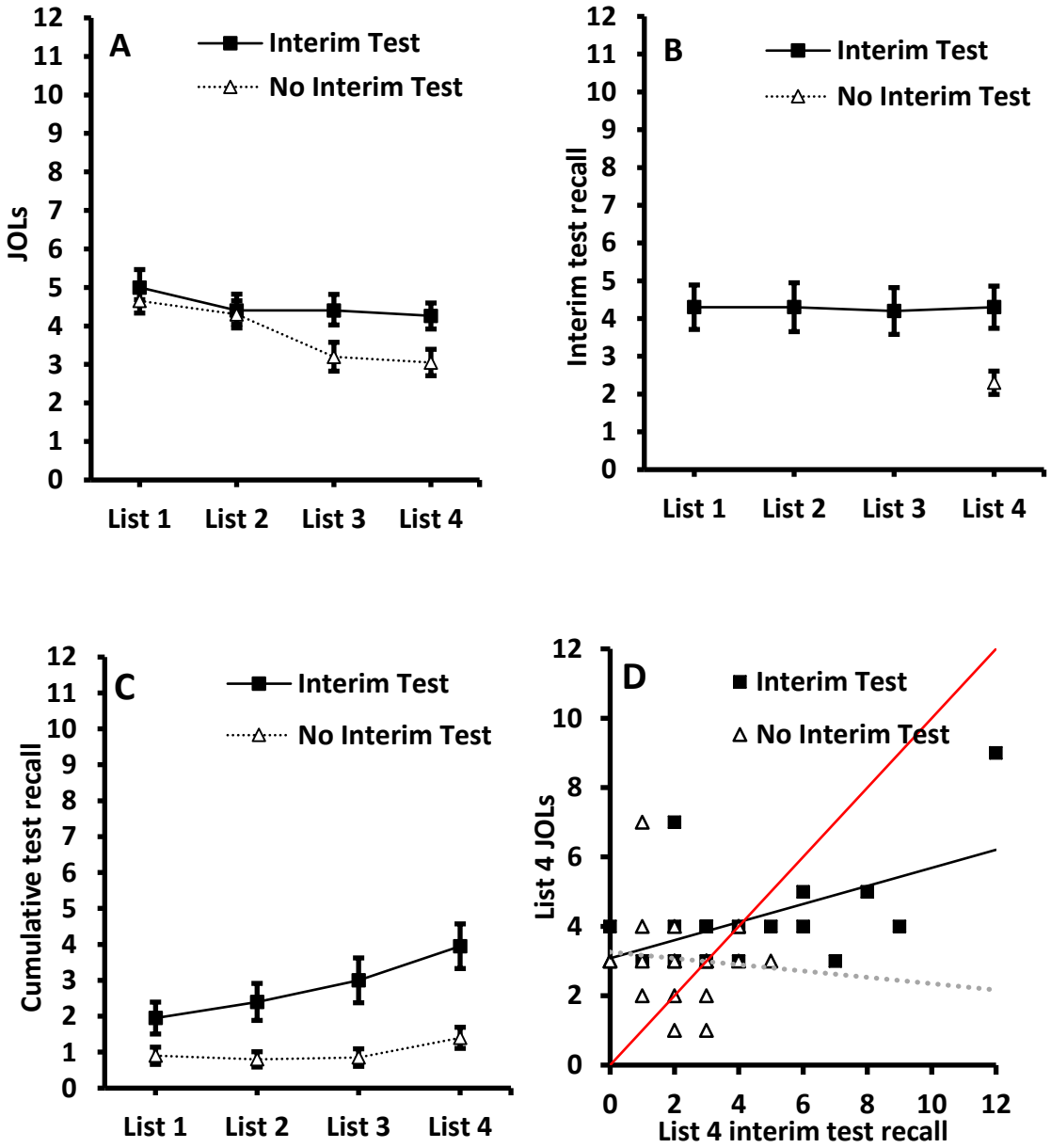

Figure 3.1: Experiment 3. Panel A: Mean JOLs across four face-name lists. Panel B: Interim test recall across four lists. Panel C: Cumulative test recall across four lists. Panel D: Calibration curve of List 4 JOLs; the red line represents perfect agreement between JOLs and recall and the black and dashed grey lines represent the relationships between JOLs and recall for the Interim Test and No Interim Test groups, respectively. Error bars represent ± 1 standard error.

participants in the No Interim Test group, mean difference = 2.00 names, 95% CI = [0.53, 3.47], Cohen's *d* = 0.87.

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 names, 95% CI = [-0.78, 3.68], Cohen's *d* = 0.42. To measure PI (mistakenly recalling another face's name from a prior list), an independent-samples *t* test was conducted. 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* = 0.85, *SD* = 1.23), mean difference = 2.05 names, 95% CI = [0.90, 3.20], Cohen's *d* = 1.14. 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 = -0.06 names, 95% CI = [-2.39, 1.19], Cohen's *d* = -0.22). 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 names, 95% CI = [1.94, 7.76], Cohen's *d* = 1.17 (see Figure 3.1C). This can be attributed to some combination of a standard backward testing effect and more attention and effort 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 names, 95% CI = [1.16, 3.94], Cohen's *d* = 1.17.

Calibration of List 4 JOLs

There are two methods to demonstrate the calibration of JOLs (i.e., the agreement between JOLs and recall performance). The first is to plot a calibration curve and the second is to calculate the absolute agreement score between JOLs and recall performance (see below for illustrations). Because the No Interim Test group only took an interim test on List 4, calibration analyses were restricted to List 4 JOLs and List 4 interim test performance.

Figure 3.1D is a calibration curve depicting the agreement between List 4 JOLs and List 4 interim test recall for both groups. The red line represents perfect calibration. Points located above the line represent overconfident participants (i.e., JOLs > recall) while ones located below the red line represent underconfident participants (i.e., JOLs < recall).

Table 2.1 depicts the numbers of underconfident, perfectly-accurate, and overconfident participants in each group. A chi-square test showed no difference in the proportions of participants in these three categories between groups, $\chi^2(2) = 1.69$, $p = 0.43$, indicating no difference in tendency toward underconfidence or overconfidence between groups.

The second method is to calculate the absolute agreement scores between JOLs and recall performance. The calibration score for each participant was calculated using the following formula:

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

Calibration scores range from 0 to 100. 0 means completely inaccurate, and 100 means

Table 2.1: The number of participants classified as underconfident, perfectly-accurate, or overconfident in Experiments 3 and 4.

Groups	Underconfident (JOL < Recall)	Perfectly-accurate (JOL = recall)	Overconfident (JOL > Recall)
Experiment 3			
No Interim Test	5	6	9
Interim Test	8	3	9
Experiment 4			
No Interim Test	10	0	10
Interim Test	14	0	6

perfectly accurate. There was no significant difference in calibration scores 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].[2]

Correlations between JOLs and test recall

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.

[2] For calibration scores, one outlier (± 2.5 *SD* from the mean) was detected in the No Interim Test group and two in the Interim Test group. Excluding these outliers yielded the same pattern: mean difference = 2.88 [-3.61, 9.36].

At the participant level, for each group, we calculated a Pearson correlation between List 4 JOLs and interim test recall for that list. Figure 3.1D depicts the correlations between List 4 JOLs and List 4 interim test recall across participants. For the No Interim Test group, there was no significant correlation, r = -.10, p = .68 (see the dashed grey line in Figure 3.1D). For the Interim Test group, there was a significantly positive correlation, r = .52, p = .02 (see the black line in Figure 3.1D). However, two outliers (± 2.5 *SD* from the regression line) were detected in the Interim Test group. Excluding the outliers yielded a smaller correlation value, r = .38, p = .12. The difference in correlations between the groups was significant including outliers, z = -1.97, p = .05, but non-significant excluding outliers, z = -1.46, p = .07. Collapsed across groups to increase power, the correlation between List 4 JOLs and interim test recall was statistically significant, r = .44, p < .01 (one outlier was detected and excluding it yielded a similar result, r = .45, p < .01).

Given that the correlations for each group were at the weak-medium level and that the sample sizes (20 in each group) were relatively small, no firm conclusions can be drawn from these results (Button et al., 2013) and hence they will not be discussed further.

Summary

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 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 Interim Test group suffered less PI in the List 4 interim test than the No Interim Test group, which again supports the involvement of factors in the retrieval processes in the forward testing effect.

Experiment 4

To generalize and conceptually replicate the findings of Experiment 3, Experiment 4 employed five 18-word lists as materials, as did Szpunar et al. (2008). 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 and they 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 at http://websites.psychology.uwa.edu.au/school/MRCDatabase/uwa_mrc.htm). Letter length was controlled between four and eight, Kucera-Francis written frequency between 100 and 850, and concreteness and familiarity between 250 and 650. These nouns were randomly divided into five lists of 18 items each. List order was counterbalanced across participants by a Latin square design: four participants in each group studied these lists in each of five 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 five 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 2.1).

At the encoding stage, each word was presented for 2 sec for participants to study. 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 will not 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, recalling words from the just-studied list and typing their answers into a blank box on screen, 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 five lists. The experiment lasted about 25 min.

Results

JOLs

Average JOLs on each of Lists 1-5 for both groups are shown in Figure 3.2A. These data were analyzed 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, $F(1, 38) = .49$, $p = .49$. 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,

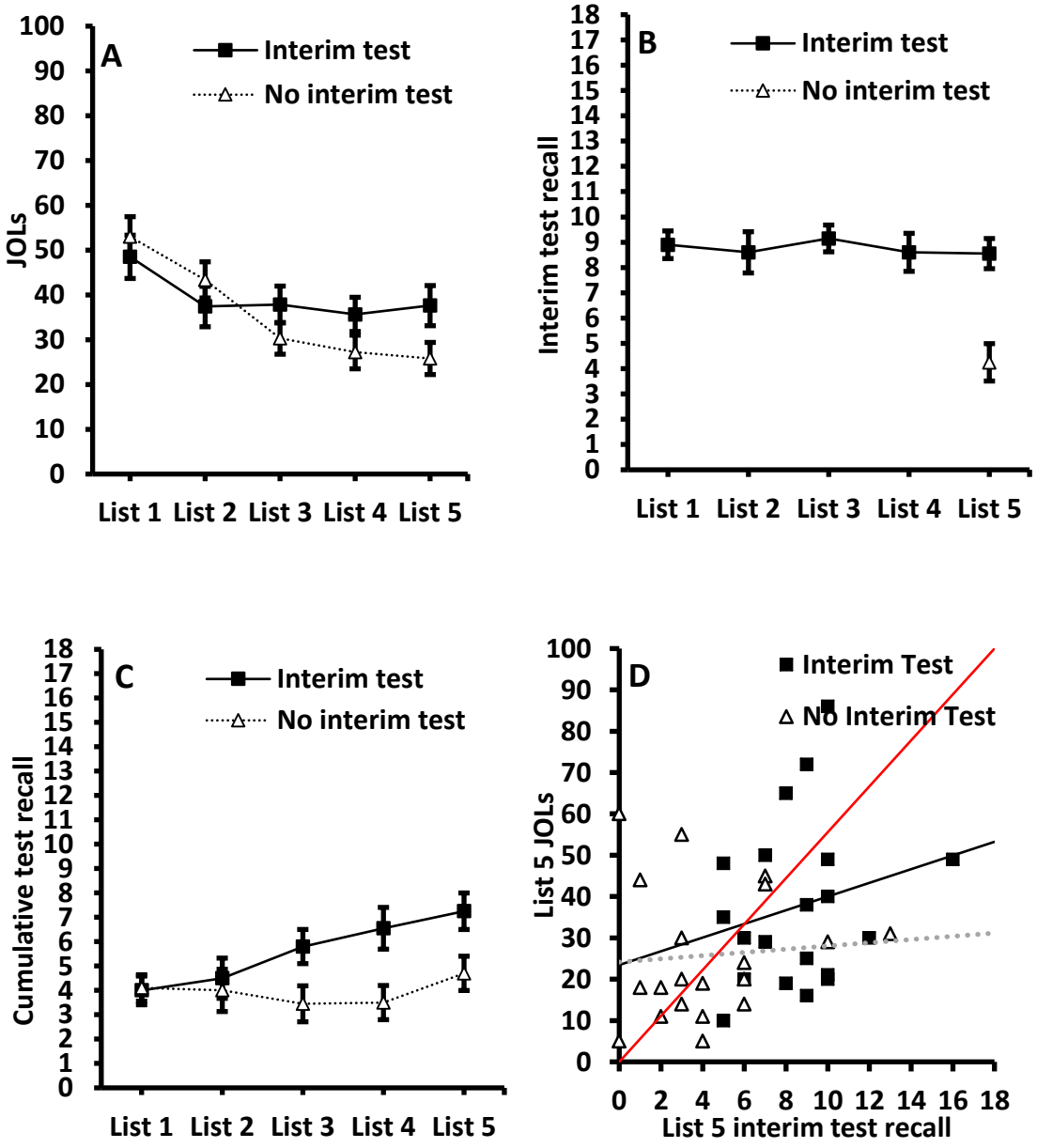

Figure 3.2: Experiment 4. Panel A: Mean JOLs across five lists. Panel B: Interim test recall across five lists. Panel C: Cumulative test recall across five lists. Panel D: Calibration curve of List 5 JOLs; the red line represents perfect agreement between JOLs and recall and the black and dashed grey lines represent the relationships between JOLs and recall for the Interim Test and No Interim Test groups, respectively. Error bars represent ± 1 standard error.

$F(1, 19) = 7.88$, $p = .01$, $\eta_p^2 = .29$.[3]

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%, 95% CI = [0.18, 23.42], Cohen's $d = 0.65$. No statistically significant difference in JOLs was detected on Lists 1-4, $-1.57 < t\text{s} < .68$, $.13 < p\text{s} < .50$.

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 3.2B. For the Interim Test group, a mixed ANOVA with List as a within-subjects variable showed that there was no main effect of List, $F(4, 76) = .24$, $p = .92$, 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 words, 95% CI = [2.38, 6.22], Cohen's $d = 1.43$. In this experiment, intrusions can only be from prior lists; current list intrusions are not meaningful because the test format was free recall. Participants in the No Interim Test group ($M = 2.30$, $SD = 4.27$) 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 = 2.05 words, 95% CI = [1.04, 5.06], Cohen's $d = 1.02$.

Cumulative test recall

[3] The quadratic trends of JOLs were also significant for both groups: No Interim Test group, $F(1, 19) = 16.39$, $p < .01$, $\eta_p^2 = .46$; Interim Test group, $F(1, 19) = 6.31$, $p = .02$, $\eta_p^2 = .25$.

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 = [-0.02, 11.62], Cohen's *d* = 0.64 (see Figure 3.2C). 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 words, 95% CI = [0.47, 4.63], Cohen's *d* = 0.78.

Calibration of List 5 JOLs

Figure 3.2D depicts the calibration curve of List 5 JOLs for both groups. Table 2.1 depicts the numbers of underconfident, perfectly-accurate, and overconfident participants in each group. A chi-square test found no difference in the proportions of participants in the three (underconfident/perfect-accurate/overconfident) categories between groups, χ^2 (1) = 0.94, p = 0.33, indicating no difference in tendency toward under/overconfidence between groups.[4]

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 scores between the groups (No Interim Test group: *M* = 83.39, *SD* = 16.19; Interim Test group: *M* = 78.94, *SD* = 11.25), mean difference = -4.44, [-4.81, 13.37].

Correlations between JOLs and test recall

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

[4] As shown in Table 2.1, no JOLs were perfectly accurate in Experiment 4, and therefore the perfectly-accurate category was excluded from the chi-square test.

Pearson correlation between JOLs and interim test recall on List 5 at the participant level. The correlations for both groups were statistically non-significant: No Interim Test group: r = .08, p = .74 (see the dashed grey line in Figure 3.2D); Interim Test group: r = .22, p =.74 (see the black line in Figure 3.2D). There was no significant difference in correlations between the groups, z = 0.42, p = .67. By collapsing the data of JOLs and interim test recall on List 5 across two groups, we observed a marginally significant Pearson correlation, r = .30, p = .06.

Summary

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.

Discussion

In both experiments, the forward testing effect was replicated when the encoding procedure was experimenter-paced. In the final list interim test, participants in the Interim Test group outperformed those in the No Interim Test group. More importantly, in both experiments, although participants' JOLs decreased across lists in both groups, JOLs in the Interim Test group decreased much less across lists than those in the No Interim Test group.

In Experiment 4, in the final list interim test, participants in the No Interim Test group experienced about 9.60 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 effect (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 realized that their minds had wandered more and more across lists (Szpunar et al., 2013) and that they exerted less and less encoding effort (as found in 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.

These results can be interpreted as 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 effect 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. 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 might 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 tests. 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 conclusion, interim tests insulate against the build-up of PI and enhance learning and retrieval of new information when the study procedure is instructor-paced. The forward testing effect is associated with metacognitive insight. Future research should explore whether people's metacognitive insight into the forward testing effect is explicit or implicit. In the next chapter, we ask whether the forward testing effect and its associated metacognitive insights generalize to inductive learning.

CHAPTER FOUR: THE FORWARD TESTING EFFECT ON INDUCTIVE LEARNING

Induction refers to the process in which people generalize their previous experience when making uncertain inferences about the environment that go beyond direct experience. Inductive learning is of considerable practical and theoretical interest for learners, educators, and researchers as it is an essential component of how individuals learn and understand the world (Holland, Holyoak, Nisbett, & Thagard, 1989). A substantial body of research has investigated how to improve inductive learning (Djonlagic et al., 2009; Giguere & Love, 2013; Kornell & Bjork, 2008a; Mathy & Feldman, 2009; Pashler & Mozer, 2013). However, it is surprising that little research has investigated how to employ testing to enhance inductive learning (Jacoby, Wahlheim, & Coane, 2010), bearing in mind that in the last 100 years, scores of experiments have revealed that repeated testing of studied information enhances its retention more effectively than restudying (Karpicke & Roediger, 2008; Roediger & Karpicke, 2006a). This chapter explores whether the forward testing effect generalizes to inductive learning.

The benefits of testing may be limited to low-level learning (e.g., facts, skills) but not extend to high-level learning (e.g., inductive learning). It is possible for example that retrieval practice focuses individuals' attention on remembering the details of exemplars, to the benefit of retention of these exemplars but to the detriment of abstraction of common characteristics shared by exemplars. However, a previous study found that repeated testing on studied categories facilitates their inductive learning. Jacoby et al. (2010) asked participants to study various bird families. In a repeated study condition, a set of exemplars and bird family names were presented for participants to study four times. In a repeated testing condition, exemplars and bird family names were shown together for participants to study once, and then they classified these exemplars three times followed by corrective feedback. In a cumulative test,

the repeatedly tested families were better classified than the repeatedly studied ones. This study revealed a clear backward testing effect on inductive learning (i.e., testing of previously studied categories enhances their inductive learning and classification).

Research investigating the forward testing effect is largely restricted to low-level learning (Pastötter & Bäuml, 2014; for detailed discussion, see Chapter 1). Specifically, previous studies documented that participants learned the target items much better if they had been tested rather than untested on previous items. Those studies show that item-level learning is susceptible to enhancement induced by interim testing. It is unknown whether testing can have a facilitatory forward effect on inductive learning. Schacter and Szpunar (2015) suggested that "An important question is whether interpolated retrieval/testing also enhances learning at a conceptual level" (p. 67). This chapter aims to answer this question.

What would we expect to happen if we evaluate the forward testing effect on category induction rather than item learning? There are reasons to predict that category learning will be less enhanced and indeed might even be unaffected by interim testing. Evidence shows that enhancing the encoding of individual exemplars can sometimes have little benefit for category learning. Category induction and stimulus distinctiveness can interact, with induction benefitting much less than identification learning (i.e., item memory) as the stimuli are rendered more distinctive (Love, 2000). For example, Smith, Redford, Washburn, and Taglialatela (2005) studied airport security screeners' ability to detect threatening items in x-ray images. A manipulation which boosted identification of specific items had virtually no effect on generalization to novel exemplars. These findings complement the many other variables known to have divergent effects on exemplar versus category learning, the best-known being the effects of amnesia resulting from temporal lobe damage. Many studies have shown that individuals with amnesia are much more impaired at item memory (recognition) than category induction (Knowlton & Squire, 1993; Reed, Squire, Patalano, Smith, &

Jonides, 1999). Theories of category induction have been successful at accounting for these interactions in terms of either the involvement of multiple independent systems underlying the two forms of learning (Ashby, Alfonso-Reese, Turken, & Waldron, 1998) or in terms of the differential demands that induction and identification place on the ability to distinguish stimulus representations (e.g., Nosofsky, Denton, Zaki, Murphy-Knudsen, & Unverzagt, 2012).

Thus, the existence of a forward testing effect on item-level learning does not imply that a parallel effect on category learning will be observed, and indeed the theoretical analysis of category learning provides at least some grounds for expecting divergent effects of testing. Experiments 5 and 6 were designed to explore whether interim testing enhances inductive learning of new categories more effectively than no interim testing or studying additional category exemplars.

Experiment 5

Method

Participants

Experiments 1-4 observed the effect sizes (i.e., Cohen's *d*s) of the forward testing effect ranging from 0.87 to 1.43. Sample size calculation was conducted using G*power (Faul et al., 2007), which showed that about 10-23 participants in each group were required to observe a significant (α = .05) forward testing effect at 0.8 power. Forty participants, 31 females, with an average age of 21.45 (*SD* = 4.42) years, were recruited from the UCL participant pool and were randomly divided into two equal-sized groups (Interim Test/Interim Math). They received £4 or course credit as compensation for participating.

Materials

The principal stimuli used were 20 paintings by each of eight to-be-studied Renaissance artists [Lucas Cranach the Elder, Andrea del Sarto, Sandro Botticelli, Paolo Veronese, Raffaello Sanzio da Urbino (known as Raphael), Jacopo da Pontormo (known as Pontormo), Cosimo Tura, and Jan van Eyck], plus 4 paintings by each of 5 filler artists [Fra Angelico, Tiziano Vecelli (known as Titian), Leonardo da Vinci, Giovanni Bellini, and Tintoretto]. These paintings were trimmed and resized to fit into a 24 × 18 cm rectangle (visual angle: about 16 º × 12 º). These artists, except Tintoretto, were divided into four sets. Each set consisted of two to-be-studied artists and one filler artist: Set 1: Cranach the Elder, del Sarto, and Angelico; Set 2: Botticelli, Veronese, and Titian; Set 3: Raphael, Pontormo, and da Vinci; Set 4: Tura, van Eyck, and Bellini. The set order was counterbalanced by using a Latin square design across participants.

Design and procedure

The experiment employed a 2 (Interim task: Interim Test/Interim Math) × 4 (List: 1-4) mixed design, with Interim task as a between-subjects variable and List as a within-subjects variable. Participants were instructed to study the painting styles of various famous painters in anticipation of a cumulative test. They were told that in the first part they would see a list of paintings, consisting of 12 paintings by each of two to-be-studied artists from one set. Following these 24 paintings, they would solve as many math problems as they could in 30 sec (e.g., *47+32 = ?*), and then the computer would decide at random whether or not to give them a short test. If it did, then 12 new paintings (four by each of these two studied artists plus another four by a different artist) would be shown one at a time in a random order and their task was to decide which artist was responsible for each painting. If it did not, they would continue solving math problems for another 60 sec. Then they would go on to the second part identical to the first except that they would learn the styles of two new artists. In

fact, the Interim Test group was tested on every list while the Interim Math group was only tested on List 4 (see the design schema in Fig. 4.1).

At the encoding phase, a painting was shown for 5 sec with the artist's last name displayed below. Paintings from the two artists were alternated in a random order in the following sequence: A1, B1, A2, B2…A12, B12 (Kornell & Bjork, 2008a). At the interim test phase, 12 new paintings were randomly presented one at a time, with the two studied artists' names and *None of these* displayed below against the option labels A-C. Participants had unlimited time to classify each painting.

Following the completion of List 4, participants were instructed to undertake a cumulative test in which 36 new paintings (four by each of the eight studied artists plus another four by Tintoretto) were shown in a random order, with eight artists' names and *None of these* displayed below against the option labels A-I. For each painting participants guessed which artist was responsible. There was no feedback in interim and cumulative tests. The experiment lasted about 35 min.

Results

Interim test performance

Figure 4.2A shows interim test accuracy. The Interim Test group correctly classified about 65% of paintings (all paintings including those from studied and new artists) and their classification accuracy did not fluctuate across lists, $F(3, 57) = 0.06$, $p = .98$, $\eta_p^2 < .01$. The Interim Test group correctly classified more List 4 paintings than the Interim Math group, difference = 2.15 paintings, 95% confidence interval (CI) = [0.77, 3.53], Cohen's $d = 1.03$, which reveals a substantial forward testing effect.

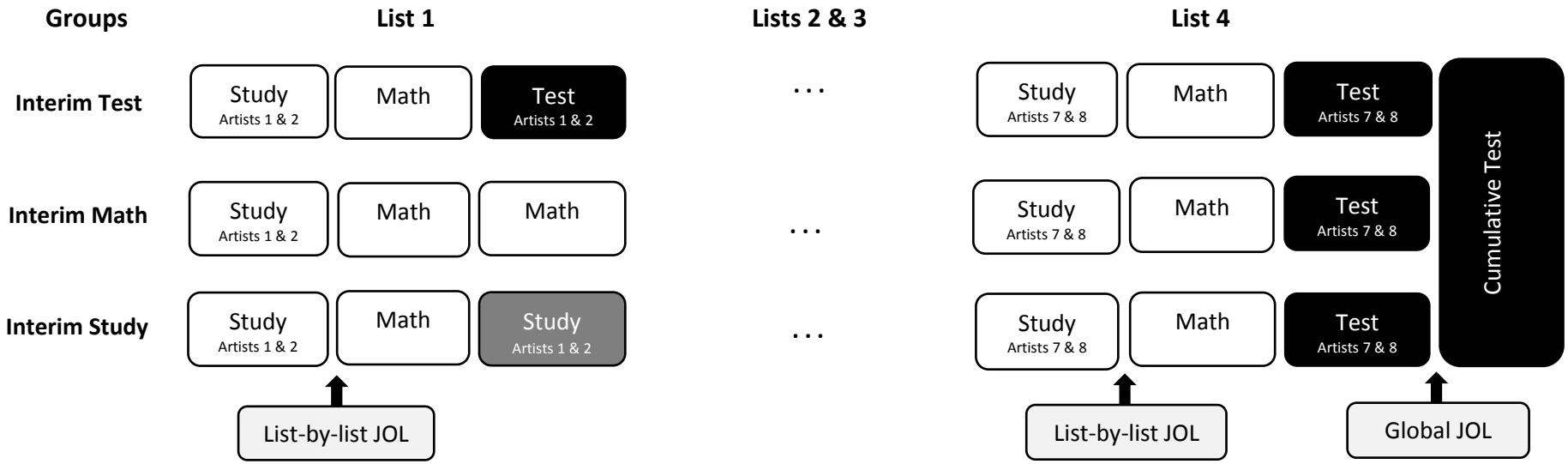

Figure 4.1: Experimental design schema for the Interim Test, Interim Math, and Interim Study groups. Lists 2 and 3 were identical to List 1 in each group, but with different artists. The Interim Study group was not included in Experiment 5. Judgments of learning (JOLs; i.e., metacognitive judgments about the degree of mastery of the studied artists' styles) were only made in Experiment 6.

Cumulative test performance

For the cumulative test, classification of Lists 1-3, List 4, and new artist's paintings were analyzed separately, as the Interim Test group underwent an interim test on each of Lists 1-3 but the Interim Math group did not, whereas both groups undertook an interim test on List 4. In the cumulative test, for List 1-3 artists, the Interim Test group correctly classified more paintings than the Interim Math group, difference = 3.35 paintings, 95% CI = [1.19, 5.51], Cohen's *d* = 0.99 (see Figure 4.2B). In addition, the Interim Test group correctly classified more List 4 artists' paintings than the Interim Math group, difference = 1.15 paintings, 95% CI = [0.36, 1.94], Cohen's *d* = 0.96, corroborating the pattern found in the List 4 interim test. For new artists, no statistically significant difference was detected between the groups, difference = -0.10 paintings, 95% CI = [-0.83, 0.63], Cohen's *d* = 0.09. The Interim Math group chose *None of these* somewhat more frequently (*M* = 8.10, SD = 7.26)

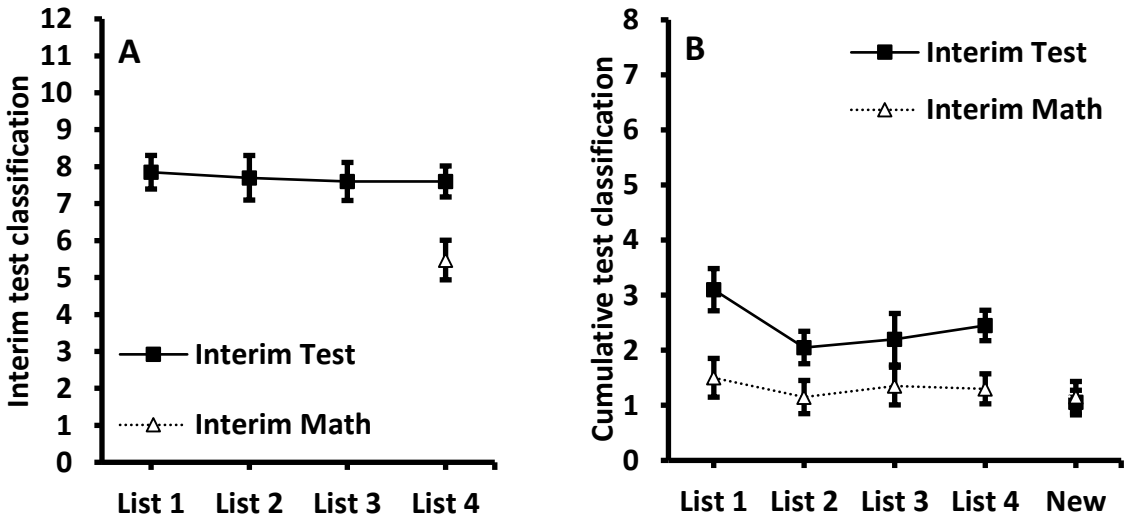

Figure 4.2: Experiment 5. Panel A: Interim test classification accuracy (no. correct) across lists. Panel B: Cumulative test classification accuracy (no. correct) across lists and for one new artist (*None of these*). Error bars represent ± 1 standard error.

than the Interim Test group (M = 6.45, SD = 3.83), although the difference was not statistically significant, difference = 1.65 paintings, 95% CI = [-2.07, 5.37], Cohen's d = 0.34.

Summary

List 4 interim test classification reveals for the first time that testing of previously studied concepts improves inductive learning of new concepts – a forward testing effect on inductive learning. In addition, the Interim Test group correctly classified nearly twice as many studied artists' paintings as the Interim Math group in the cumulative test, indicating that interim testing enhances inductive learning more effectively than no interim testing.

Experiment 6

Experiment 6 introduced four modifications. In Experiment 5's cumulative test, List 1 artists were always presented as options A and B, List 2 artists as options C and D, List 3 artists as options E and F, List 4 artists as options G and H, and *None of these* (new artist) as option I. This consistent placement of the response options might have aided responding. In Experiment 6, the placement of response options was therefore randomized. In Experiment 5, the Interim Test group was exposed to more paintings than the Interim Math group, because 12 new paintings were presented in each interim test. The second change in Experiment 6, therefore, was to include an Interim Study group. Participants in this group studied 12 new paintings (four from each of two studied artists plus another four from a different artist – the same pictures that were shown in the corresponding test for the Interim Test group) following each of Lists 1-3 and were tested on List 4.

In Experiment 5's cumulative test, there was no difference in classification accuracy for paintings by new artists. There were only four such paintings, and hence it is difficult to explore participants' discrimination between studied and new artists. Therefore, the third change in Experiment 6 was that we added four more paintings by a new artist in the

cumulative test. Finally, following study of each list, participants were asked to make a judgment of learning (JOL) by typing in a number (1-9) to indicate their mastery of the two artists' painting styles, and after the completion of List 4, participants were also asked to rate their mastery of all 8 artists' painting styles.

Method

Participants

Based on the effect size in Experiment 5 (Cohen's d = 1.03), 16 participants in each group were required to observe a significant forward testing effect at 0.9 power. Because Experiment 6 employed three groups in total, an analysis of variance (ANOVA) would be conducted to explore the differences among groups, which theoretically requires larger sample sizes than a *t*-test. Therefore, the sample size was increased to 24 participants in each group. Seventy-two participants, 45 females, with an average age of 25.44 (SD = 7.32) years, were recruited from the UCL participant pool and were randomly divided into three groups (Interim Test/Interim Math/Interim Study). They received £4 or course credit as compensation for participating.

Materials, design, and procedure

The same paintings plus another four by Jan Brueghel the Elder were employed. Experiment 2 involved a 3 (Interim task: Interim Test/Interim Math/Interim Study) × 4 (List: 1-4) mixed design. The procedure was the same as in Experiment 5 with the following exceptions. Participants were informed that, after studying 24 paintings and solving math problems for 30 sec, the computer would randomly decide the next task. If it decided to give them a short test, 12 new paintings would be presented one at a time and their task was to classify each painting. If it decided to give them more math problems, they would continue solving math problems for another 60 sec. If it decided to give them more new paintings, 12

new paintings (four by each of two studied artists and four by a different artist) would be presented with the artist's name or *None of these* displayed below, one at a time for 5 sec in a random order. In fact, the Interim Test group was tested on every list. The Interim Math group continued solving math problems following each of Lists 1-3 and was tested on List 4. The Interim Study group studied 12 new paintings following each of Lists 1-3 and was tested on List 4 (see the design schema in Figure 4.1).

Immediately following studying of each list, participants answered the question "How well do you think you learned the two studied artists' painting styles?" by typing in a number ranging from 1 (not very well) to 9 (very well). Following the completion of List 4, participants answered the question "How well do you think you learned the eight studied artists' painting styles?" with the same response scale. Then all participants undertook a cumulative test, in which 40 new paintings were presented in a random order. The studied artists' names were positioned against option labels A-H in a different random configuration for each participant, with *None of these* always as option I.[5] The order of studied artists' names in the cumulative test was constant across test trials. The experiment lasted about 35 min.

Results

Interim test performance

Figure 4.3A shows interim test classification. The Interim Test group's classification accuracy did not fluctuate across lists, $F(3, 69) = .22$, $p = .88$, $\eta_p^2 = .01$. For the List 4 interim test classification, a one-way ANOVA showed a main effect of Interim task, $F(2, 69) = 4.85$, $p = .01$, $\eta_p^2 = .12$. Again revealing a forward testing effect, the Interim Test group correctly

[5] Studied artists' names were randomised but the option labels were consistently in alphabetical order (i.e., A, B, C…I).

classified more List 4 paintings than the Interim Math group, difference = 2.13 paintings, 95% CI = [0.51, 3.74], Cohen's *d* = 0.78, and more than the Interim Study group, difference = 2.17 paintings, 95% CI = [0.58, 3.75], Cohen's *d* = 0.81, but there was no significant difference between the Interim Study and Interim Math groups, difference = -0.04 paintings, 95% CI = [-1.64, 1.56], Cohen's *d* = 0.02.

Cumulative test performance

Figure 4.3B shows cumulative test classification. For List 1-3 artists, a one-way ANOVA revealed a main effect of Interim task, $F(2, 69) = 9.35$, $p < .001$, $\eta_p^2 = .21$. The Interim Test group correctly classified more paintings than the Interim Math group, difference = 3.42 paintings, 95% CI = [1.53, 5.30], Cohen's *d* = 1.08, and more than the Interim Study group, difference = 3.08 paintings, 95% CI = [1.19, 4.98], Cohen's *d* = 0.97, but there was no significant difference between the Interim Study and Interim Math groups, difference = 0.33 paintings, 95% CI = [-1.11, 1.78], Cohen's *d* = 0.13. For the List 4 artists, a one-way ANOVA yielded a main effect of Interim task, $F(2, 69) = 3.81$, $p = .03$, $\eta_p^2 = .10$. The Interim Test group correctly classified more paintings than the Interim Math group, difference = 1.04 paintings, 95% CI = [0.23, 1.86], Cohen's *d* = .76, and more than the Interim Study group, difference = 1.08 paintings, 95% CI = [0.20, 1.96], Cohen's *d* = 0.73, but there was no significant difference between the Interim Study and Interim Math groups, difference = -0.04 paintings, 95% CI = [-0.79, 0.71], Cohen's *d* = 0.03. These results corroborate the pattern in the List 4 interim test.

For new artists, a one-way ANOVA showed a main effect of Interim task, $F(2, 69) = 3.21$, $p < .05$, $\eta_p^2 = .09$. The Interim Test group correctly classified more new artists' paintings than the Interim Math group, difference = 1.13 paintings, 95% CI = [0.01, 2.24], Cohen's *d* = 0.60, and more than the Interim Study group, difference = 1.33 paintings, 95%

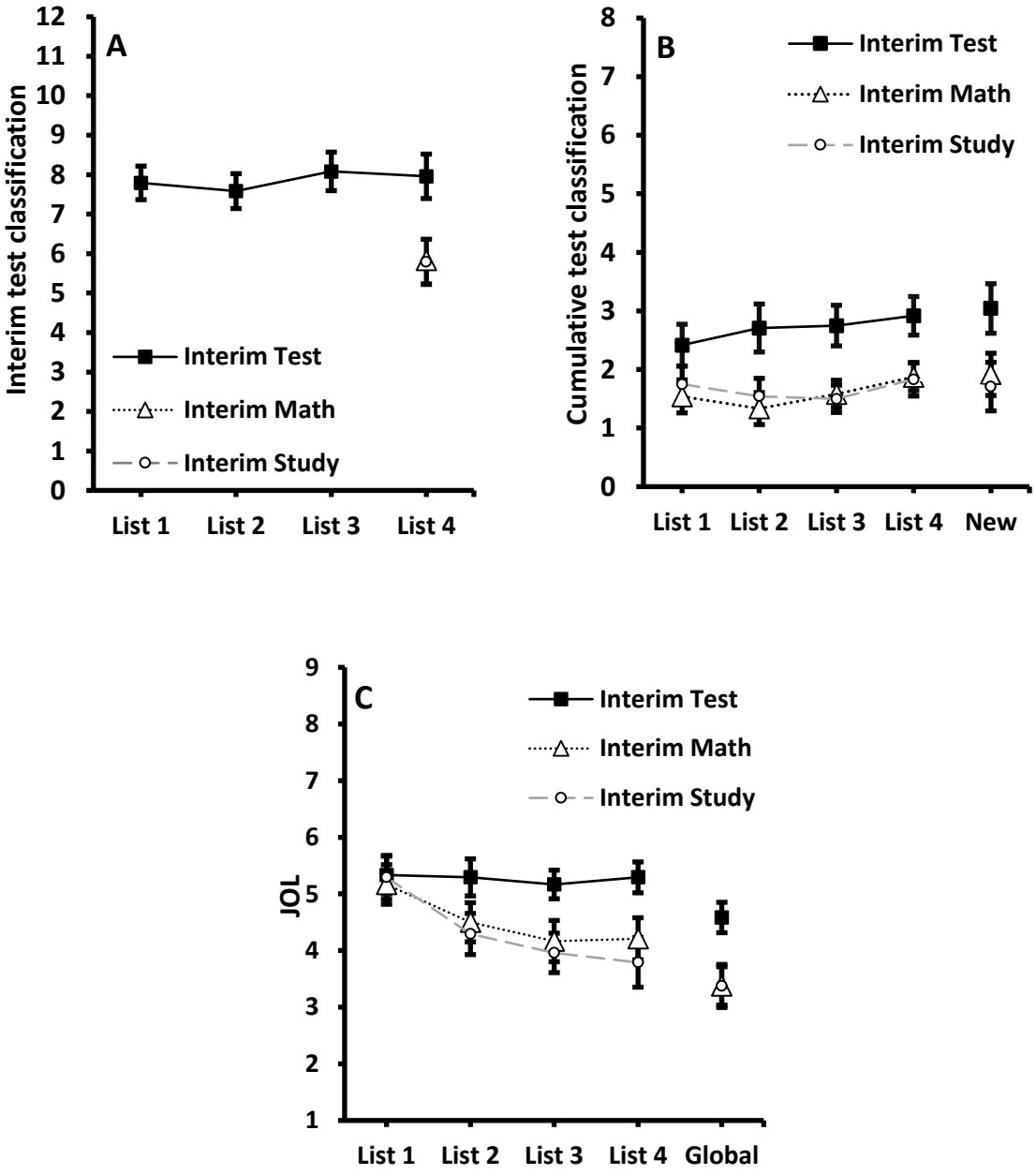

Figure 4.3: Experiment 6. Panel A: Interim test classification accuracy (no. correct) across lists. Panel B: Cumulative test classification accuracy (no. correct) across lists and for two new artists (*None of these*). Panel C: List-by-list and global JOLs. Error bars represent $\pm$ 1 standard error.

CI = [0.14, 2.53], Cohen's d = 0.66, but there was no significant difference between the Interim Study and Interim Math groups, difference = -0.21 paintings, 95% CI = [-1.31, 0.90], Cohen's d = 0.11. These results indicate that the Interim Test group was better able to discriminate studied from new artists' paintings than the other two groups.

Figure 4.3C shows list-by-list and global JOLs. For list-by-list JOLs, a mixed ANOVA with Interim task as a between-subjects variable and List as a within-subjects variable showed that list-by-list JOLs decreased linearly across lists, $F(1, 69) = 20.02$, $p < .001$, $\eta_p^2 = .29$, and there was a main effect of Interim task, $F(2, 69) = 3.17$, $p < .05$, $\eta_p^2 = .09$. There was a linear interaction between List and Interim task, $F(2, 69) = 4.72$, $p = .01$, $\eta_p^2 = .14$. JOLs decreased linearly list-by-list in the Interim Math and Interim Study groups (Interim Math: $F(1, 23) = 7.50$, $p = .01$, $\eta_p^2 = .33$; Interim Study: $F(1, 23) = 15.62$, $p = .001$, $\eta_p^2 = .68$), but did not drop across lists in the Interim Test group, $F(3, 69) = .10$, $p = .96$, $\eta_p^2 < .01$. These results reveal that the Interim Math and Interim Study groups realized the waning of their learning across lists. In contrast, the Interim Test group correctly recognized that the level of their inductive learning was maintained across lists.

For List 4 JOLs, a one-way ANOVA revealed a main effect of Interim task, $F(2, 69) = 4.46$, $p = .02$, $\eta_p^2 = .11$. List 4 JOLs in the Interim Test group were higher than those in the Interim Math group, difference = 1.08, 95% CI = [0.16, 2.01], Cohen's d = 0.69, and higher than those in the Interim Study group, difference = 1.50, 95% CI = [0.46, 2.54], Cohen's d = 0.86, but there was no significant difference between the Interim Study and Interim Math groups, difference = -0.42, 95% CI = [-1.57, 0.74], Cohen's d = 0.21. These results reveal that List 4 JOLs were aligned with List 4 interim test classification.

For global JOLs (i.e., mastery ratings for all artists' painting styles following the completion of the List 4 interim test), a one-way ANOVA showed a main effect of Interim

task, $F(2, 69) = 4.40$, $p = .02$, $\eta_p^2 = .11$. Global JOLs in the Interim Test group were higher than those in the Interim Math group, difference = 1.21, 95% CI = [0.34, 2.08], Cohen's d = 0.82, and higher than those in the Interim Study group, difference = 1.21, 95% CI = [0.27, 2.14], Cohen's d = 0.77, but there was no significant difference between the Interim Study and Interim Math groups, difference = 0.00, 95% CI = [-1.03, 1.03]. These results reveal that global JOLs aligned with cumulative test classification.

Collapsing data across groups, there was a positive correlation between List 4 JOLs and List 4 interim test classification, $r = .33$, $F(1, 71) = 8.81$, $p = .004$, $R^2 = .11$, adjusted $R^2 = .10$, and a positive correlation between global JOLs and cumulative test classification of studied artists' paintings, $r = .34$, $F(1, 71) = 9.07$, $p = .004$, $R^2 = .12$, adjusted $R^2 = .10$.

In Experiment 6, List 4 interim test classification reveals that interim testing enhances inductive learning of new categories more effectively than no interim testing or studying more new exemplars – a forward testing effect. In the cumulative test, the Interim Test group was better able to classify studied artists' paintings and discriminate between the paintings of studied and new artists than the other two groups. The Interim Math and Interim Study groups recognized the reduction in their inductive learning across lists whereas the Interim Test group was aware of the maintenance of their learning across lists.

Summary

Experiment 6 revealed that interim testing is more effective for enhancing learning of new categories than studying additional paintings or doing nothing. The forward effect of testing on inductive learning is associated with metacognitive awareness.

Discussion

Previous research has shown that many variables have stronger effects on item (exemplar) versus inductive (category) learning (Knowlton & Squire, 1993; Love, 2000).

However, contrary to the hypothesis that interim testing might have a smaller effect on inductive learning than it has on item memory (or even no facilitatory effect), both of the experiments reported here clearly reveal a robust forward testing effect, a finding which has not previously been demonstrated. A few factors may contribute to the facilitatory effect of interim testing on inductive learning.

In the absence of interim tests, inductive learning might have decreased across successive lists (Jing et al., 2016; Pastötter et al., 2011; Szpunar et al., 2013; Szpunar et al., 2008; Experiments 1-4), but interim tests maintained subsequent inductive learning of new categories. In the absence of interim tests, people's minds may wander, and less and less attention and effort is directed to learning across successive lists (Pastötter et al., 2011; Pastötter et al., 2013; Szpunar et al., 2013; Experiments 1-4), which leads to deterioration of subsequent inductive learning. Prior interim tests act as warnings of upcoming interim tests and maintain people's test expectancy at a high level (Weinstein et al., 2014). Expecting a future testing enhances subsequent learning (Szpunar et al., 2007). Cho et al. (2016) proposed that retrieval failures in prior interim tests motivate people to commit more effort to encoding subsequent new information (Kornell et al., 2009; Potts & Shanks, 2014; Experiments 1-4). Thus, in the current research, incorrect classifications on prior interim tests might encourage the Interim Test group to commit more effort to learning new categories.

Pastötter et al. (2011) proposed an encoding reset theory to explain the forward testing effect which may operate as well as or instead of the aforementioned motivational mechanisms. Pastötter et al. suggested that interim testing causes internal context changes between successive lists, which reset subsequent encoding of new information and render it as effective as encoding of prior information. Evidence for this mechanism comes from a study by Pastötter et al. (2008). These researchers measured participants' brain oscillatory activity while encoding two lists of words. Participants were instructed to either perform an

imagination task or not following encoding of the first list, and then studied the second list. Alpha (8-14 Hz) and theta (4-7 Hz) power (synchrony in brain oscillations), which are linked to reduced attention, increased across lists if participants did not perform the imagination task. The inference is that the imagination task produced an internal context change between the lists, and this context change attenuated the increase in alpha and theta power that would otherwise have occurred. Thus internal context change between lists resets the encoding of new information and makes it as effective as prior encoding.

Besides variations in the learning phase, variations in the retrieval phase might contribute to the forward testing effect on inductive learning. Cho et al. (2016) postulated that retrieval failures in prior interim tests encourage people to adopt more efficient retrieval strategies and commit more retrieval effort in subsequent interim tests. According to this proposal, the classification failures in the interim tests on each of Lists 1-3 motivated the Interim Test group to improve their classification strategies and commit more effort in the List 4 interim test.

Interim testing enhanced participants' classification of studied artists' painting styles and improved their discrimination between studied and new artists' paintings. The present research identifies two interlinked mechanisms by which this can happen. First, the forward testing effect implies that subsequent encoding of exemplars is enhanced by prior interim tests. Secondly, the act of testing exemplars in the interim tests serves to consolidate them – the backward testing effect. Jacoby et al. (2010) found that testing can enhance retention of tested exemplars. Better remembered exemplars produce a more useful source for generalization in an inductive test (Anderson, 2000; Jacoby et al., 2010; Murphy, 2002).

Correct classification also requires discrimination among different painting styles. Kornell and Bjork (2008a) found that studying different artists' paintings in an alternating

way enhances inductive learning more effectively than studying each artist's paintings blocked together. Kornell and Bjork proposed that spacing facilitates discrimination among different artists' painting styles. Similarly, interim testing may improve discrimination among different painting styles. Previous research has shown that retrieval practice leads to deeper and more elaborative learning than restudying (Pyc & Rawson, 2012). During interim tests, participants might modify their abstraction of the two studied artists' painting styles (knowledge of characteristic features shared by exemplars) and highlight the difference between the artists' styles. It has been noted that interim testing enriches list context information, which highlights list discriminability (Szpunar et al., 2008; Experiments 1-4). Interim testing might have differentiated different lists' painting styles more effectively than math problem solving or studying more new exemplars. The difference in cumulative test classification of studied artists' paintings might also be attributed to the fact that interim tests strengthened the associations between artists' names and their corresponding styles (Cho et al., 2016; Weinstein et al., 2011; Experiments 1-4).

Experiment 6 observed that, in the List 4 interim test as well as in the cumulative test, the Interim Study group failed to classify paintings any better than the Interim Math group. This might seem surprising given that the Interim Study group had the opportunity at the end of each list to study four additional paintings by the two target artists. However this lack of benefit of additional study opportunities is in line with many comparable findings in the backward testing effect (e.g., Roediger, Agarwal, McDaniel, & McDermott, 2011) and the rereading effect (e.g., Callender & McDaniel, 2009) literature. This finding serves to emphasize that the benefit of testing seen in the Interim Test group is not simply due to additional exposure to relevant learning materials. It is the act of being tested on one's knowledge that causes the benefit.

People's metamemory monitoring is sensitive to the deterioration of inductive learning across lists in the absence of interim tests. When making list-by-list JOLs, people may replay the learning process in their mind and compare it with previous learning. They may realize that their mind-wandering is increasing and their learning effort decreasing across lists (Experiments 1-4). In contrast, list-by-list JOLs do not fluctuate across lists when an interim test is administered following each list. People may realize that subsequent inductive learning is as effective as prior learning. In addition, interim test classification performance informs people of the consistency of their inductive learning across lists.

In Experiment 6, global JOLs were made following the List 4 interim test and hence might be affected by List 4 interim test classification. The Interim Test group outperformed the other two groups in the List 4 interim test, which may have induced them to report higher global JOLs than the other two groups. To test this idea, we explored the correlation between List 4 interim test classification and global JOLs at the participant level. Consistent with this idea, the correlation was positive, $r = .26$, $F(1, 71) = 5.21$, $p = .03$, $R^2 = .07$, adjusted $R^2 = .06$. Anchoring may provide another possible mechanism: the Interim Test group gave higher list-by-list JOLs to List 4 than the other two groups and these JOLs might act as anchors for global JOLs, driving higher global JOLs in the Interim Test group than in the other two groups. To test this idea, we explored the correlation between List 4 JOLs and global JOLs at the participant level. There was a positive correlation, $r = .81$, $F(1, 71) = 134.67$, $p < .001$, $R^2 = .66$, adjusted $R^2 = .65$.

In conclusion, interim testing enhances inductive learning more effectively than no interim testing or studying more new exemplars – the forward testing effect on inductive learning. This forward testing effect is associated with metacognitive awareness. The present experiments (Experiments 5 and 6) found a forward testing effect and Jacoby et al. (2010) found a backward testing effect on inductive learning. Collectively, these findings lead to a

strong recommendation that interim testing should be employed to enhance inductive learning in the classroom and elsewhere.

CHAPTER FIVE: THE TRANSFERABILITY OF THE FORWARD TESTING EFFECT

The main aim of this chapter is to explore whether interim testing of studied information from one domain can enhance learning and retention of new information from a *different* domain – the *transferability of the forward testing effect*.

Rationale of Experiments 7-9

In all previous forward testing effect studies (including Experiments 1-6 in this thesis), the type of material has always been the same across lists/segments, demonstrating that testing of studied information in one domain enhances learning and retention of new information in the same domain – the classic forward testing effect. What remains unknown is whether the forward testing effect is transferable across different domains of learning. The key difference between Experiments 7-9 and previous studies was that in the target (final) block the type of material used was different from that used in the preceding blocks, enabling the transferability of the effect to be assessed.

It is important to explore the transferability of this effect because, in natural learning situations, the types of to-be-studied material are frequently switched (Hausman & Kornell, 2014). For example, high school students may take a history class, then a geography class, and then a biology class. Even within a class, the content frequently varies. Art students, for example, may learn about the history of painting, and then about the painting styles of different artists. Besides practical implications, exploring the transferability of this effect also has theoretical implications.

Would we expect the forward testing effect to be transferable? Given that previous studies have shown that the effect is robust, intuitively we would expect an affirmative answer. In addition, the retrieval-effort theory predicts successful transfer: Recall failures in

prior interim tests should induce greater retrieval effort in the subsequent interim tests, facilitating recall performance. However, there are at least three reasons to predict no or minimal transfer.

First, a few of the aforementioned theories (i.e., the theories discussed in Chapter 1) predict no or minimal transfer. For example, the release from PI mechanism cannot contribute to transfer because switching material types prior to the final target learning stage means that minimal PI will occur. The activation-facilitation mechanism cannot contribute to transfer either because different types of material are completely unrelated. The encoding/retrieval strategy mechanisms can contribute only minimally to transfer because different types of material require different encoding and retrieval strategies.

Second, even the encoding-engagement theories (i.e., the test-expectancy, failure-encoding-effort, and encoding-reset theories) might predict minimal transfer. For example, it is unclear whether or not a no-test group's test expectancy will decrease across lists when material types are switched. Therefore, the test-expectancy theory is unable to make a clear *a priori* prediction. The failure-encoding-effort theory also fails to yield a clear prediction, because it cannot assert for certain whether retrieval failures in tests on studied information from one domain will enhance the learning of new information from a different domain. The encoding-reset mechanism can contribute little to transfer because switching material types also induces substantial context changes (e.g., Ellis & Montague, 1973; Emery et al., 2008; Lustig, May, & Hasher, 2001; Nunes & Weinstein, 2012), which will "reset" subsequent encoding regardless of whether interim tests are administered or not.

Third, even if interim testing of studied information from one domain enhances effort toward encoding new information in a different domain, enhanced encoding effort may produce null improvement in learning and recall of new information – the "*labor in vain*

effect" (Callender & McDaniel, 2009; DeLozier & Dunlosky, 2015; Nelson & Leonesio, 1988; Experiment 6 in Chapter 4) – and lead to little transferability of the forward testing effect.

In summary, there are reasons to expect that the forward testing effect will transfer. However, many theories predict no or minimal transfer. Nevertheless, in many natural situations, learning content frequently varies within and across classes (and lectures). Therefore the current chapter explores this important issue – the transferability of the forward testing effect.

Experiment 7

Experiment 7 had three aims. The first was to conceptually replicate the classic forward testing effect (i.e., testing of studied information in a domain enhances learning and retention of new information in the *same* domain). To achieve this aim, two groups (Same-Test and Same-Math) of participants were employed to study four lists of face-name pairs, with the Same-Test group tested on every list while the Same-Math group was only tested on List 4. The second aim was to explore the transferability of the forward testing effect (i.e., whether testing of studied information in a domain enhances learning and retrieval of new information in a *different* domain). To assess transfer, two other groups (Different-Test and Different-Math) of participants were employed to study three lists of Swahili-English pairs followed by a list of face-name pairs, with the Different-Test group tested on every list while the Different-Math group was only tested on List 4. The third aim was to conceptually replicate Weinstein et al.'s (2014) test expectancy findings (i.e., test expectancy increases in the test group across lists but decreases in the no-test group), and therefore all participants were asked to report their test expectancy before studying each list.

Method

Participants

In previous studies (e.g., Weinstein et al., 2011) and Experiments 2 and 3, the observed effect sizes (Cohen's *d*s) of the forward testing effect on the learning of face-name pairs ranged from 0.87 to 1.47. Using these effect sizes and the G*Power program (Faul et al., 2007), the calculated sample size is that about 8-29 participants in each group are required to observe a significant (α = .05; power = 0.90) forward testing effect on the learning of face-name pairs. The sample size was, therefore, set at 20 participants in each group. In total, 82 participants, mean age = 22.77 (*SD* = 5.90) years, including 64 females, were recruited from the University College London (UCL) participant pool.[6] No participants had previously taken part in Experiments 1-6 or other forward testing effect studies and they reported no prior experience of the Swahili language. They received course credits or payment as compensation. Participants were randomly divided into four groups, with 20 in the Same-Test, 20 in the Same-Math, 21 in the Different-Test, and 21 in the Different-Math groups.

Materials

Forty-eight male faces were taken from the FEI face database developed by Thomaz and Giraldi (2010) (available at http://fei.edu.br/~cet/facedatabase.html). Forty-eight male names were taken from Baby Centre UK (available at http://www.babycentre.co.uk/a25017755/top-baby-boy-names-2015). Faces and names were randomly paired and face-name assignment was consistent across participants. These face-name pairs were randomly divided into four lists, each comprising 12 pairs. Thirty-six Swahili-English word pairs were obtained from Nelson and Dunlosky (1994) and were separated into three lists according to the recall probabilities in Nelson and Dunlosky (1994) to ensure roughly equivalent memorability across lists. The Same-Test and Same-Math

[6] Because of over-recruitment, there were two more participants in Experiment 7 than pre-planned. Excluding them did not affect the overall conclusions.

groups studied four lists of face-name pairs, whereas the Different-Test and Different-Math groups studied three lists of Swahili-English pairs followed by a list of face-name pairs. For the Same-Test and Same-Math groups, the order of the face-name lists was counterbalanced across participants using a Latin square design. For the Different-Test and Different-Math groups, the order of the Swahili-English lists was randomized across participants and the four lists of face-name pairs were employed in a roughly equal frequency (about five times) as the fourth list.

Design and procedure

The experiment adopted a 2 (Material: same/different) × 2 (Interim task: test/math) between-subjects design. The Same-Test and Same-Math groups were instructed to study four lists of face-name pairs whereas the Different-Test and Different-Math groups were told to study three lists of Swahili-English pairs and then a list of face-name pairs. All four groups were warned at the outset of the cumulative test, in which all to-be-studied materials would be tested. They were also told that the computer would randomly decide whether to give them a short test or more math problems after studying each list and solving some math problems. In fact, the Same-Test and Different-Test groups were tested on every list while the Same-Math and Different-Math groups were only tested on List 4 (see Figure 5.1).

Before studying each list, participants were asked to report whether they thought they would be tested on the subsequent list by dragging and clicking a pointer on a scale ranging from 0 ("*I am sure there will not be a test*") to 100 ("*I am sure there will be a test*"). In each list's study phase, 12 face-name pairs or 12 Swahili-English pairs were randomly presented one by one, for 5 sec each, with faces or Swahili words on the left side and names or English words on the right side of the screen. Following each list, a distractor task was administered: participants were instructed to solve some math problems (e.g. *63+18= ?*) for 60 sec. After that, participants either took a short interim test on the just-studied list or continued solving

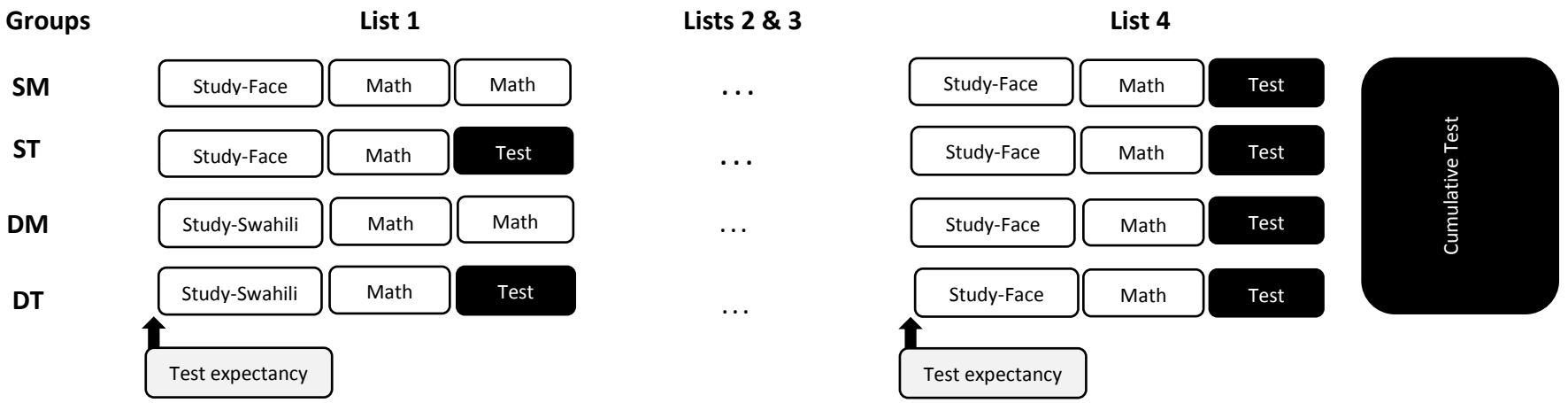

Figure 5.1: Experiment 7. The Same-Test (ST) and Same-Math (SM) groups studied four lists of face-name pairs while the Different-Test (DT) and Different-Math (DM) groups studied three lists of Swahili-English pairs followed by a list of face-name pairs. Prior to studying each list, all four groups reported their test expectancy. The Same-Test and Different-test groups took interim tests on all four lists whereas the Same-Math and Different-Math groups only took an interim test on List 4. All four groups took a cumulative test.

math problems for another 60 sec. In the interim tests, the faces or Swahili words were presented one by one in a new random order and participants were asked to recall the names or English translations. No feedback was given.

Following the completion of List 4, all participants took a cumulative test. For the Same-Test and Same-Math groups, all 48 faces were presented one by one in a random order. For the Different-Test and Different-Math groups, all 36 Swahili words were presented one by one in a random order and then the 12 faces were presented one by one in a random order. As in the interim tests, there was no feedback in the cumulative test.

In both the study and (interim and cumulative) test phases, prior to each study or test trial, a cross sign was presented for 0.5 sec to mark the interstimulus interval (ISI). Participants completed the interim and cumulative tests in their own time, and they were allowed to leave some questions blank if they did not remember the answers. The experiment lasted about 30 min.

Results

Close misspellings were counted as correct following Weinstein et al. (2011). For example, both "Toney" and "tony" were accepted as correct if the correct answer was "Tony". Two assessors independently scored the recall performance. 98.8% of their scores were in agreement and the discrepant scores were settled through a discussion.

List 1-3 interim test recall

Table 5.1 depicts the Same-Test and Different-Test groups' correct recall in each of the List 1-3 interim tests. A mixed analysis of variance (ANOVA), with Material (same/different) as a between-subjects variable and List (1-3) as a within-subjects variable, revealed that interim test recall did not fluctuate across lists, $F(2,78) = 2.59$, $p = .08$, $\eta_p^2 = .06$, and there was no interaction between Material and List, $F(2,78) = .35$, $p = .70$, η_p^2

= .01. The Different-Test group significantly outperformed the Same-Test group, $F(1,39) = 17.36$, $p < .001$, $\eta_p^2 = .31$. As can be seen in Table 5.1, the face-name pairs were more difficult to remember than the Swahili-English pairs.

Table 5.1: Mean (*SD*) List 1-3 interim test recall in Experiment 7.

Groups	List 1	List 2	List 3
Same-Test	4.05 (2.14)	4.25 (2.36)	3.70 (1.95)
Different-Test	6.57 (2.99)	7.14 (2.35)	5.95 (2.56)

List 4 interim test recall

Figure 5.2A depicts List 4 interim test recall. An ANOVA, with Material and Interim task as between-subjects variables, revealed a main effect of Interim task, $F(1,78) = 14.89$, $p < .001$, $\eta_p^2 = .16$, indicating that interim testing, compared to no interim testing (solving math problems), enhanced learning and retrieval of new information. There was a main effect of Material, $F(1,78) = 6.98$, $p = .01$, $\eta_p^2 = .08$, reflecting the fact that a switch of material types enhanced recall. There was no interaction between Interim task and Material, $F(1,78) = .02$, $p = .88$, $\eta_p^2 < .001$.

An independent-samples *t* test showed that the Same-Test group ($M = 3.80$, $SD = 2.40$) significantly outperformed the Same-Math group ($M = 2.10$, $SD = 2.40$), difference = 1.85 names, 95% CI = [0.43, 3.27], Cohen's $d = 0.84$, indicating that interim testing of studied information from one domain enhances learning and retrieval of new information from the *same* domain – the classic forward testing effect. Similarly, the Different-Test group ($M = 4.95$, $SD = 1.88$) outperformed the Different-Math group ($M = 3.24$, $SD = 2.05$), difference = 1.71 names, 95% CI = [0.49, 2.94], Cohen's $d = 0.87$, revealing that interim testing of studied information from one domain enhances learning and retrieval of new

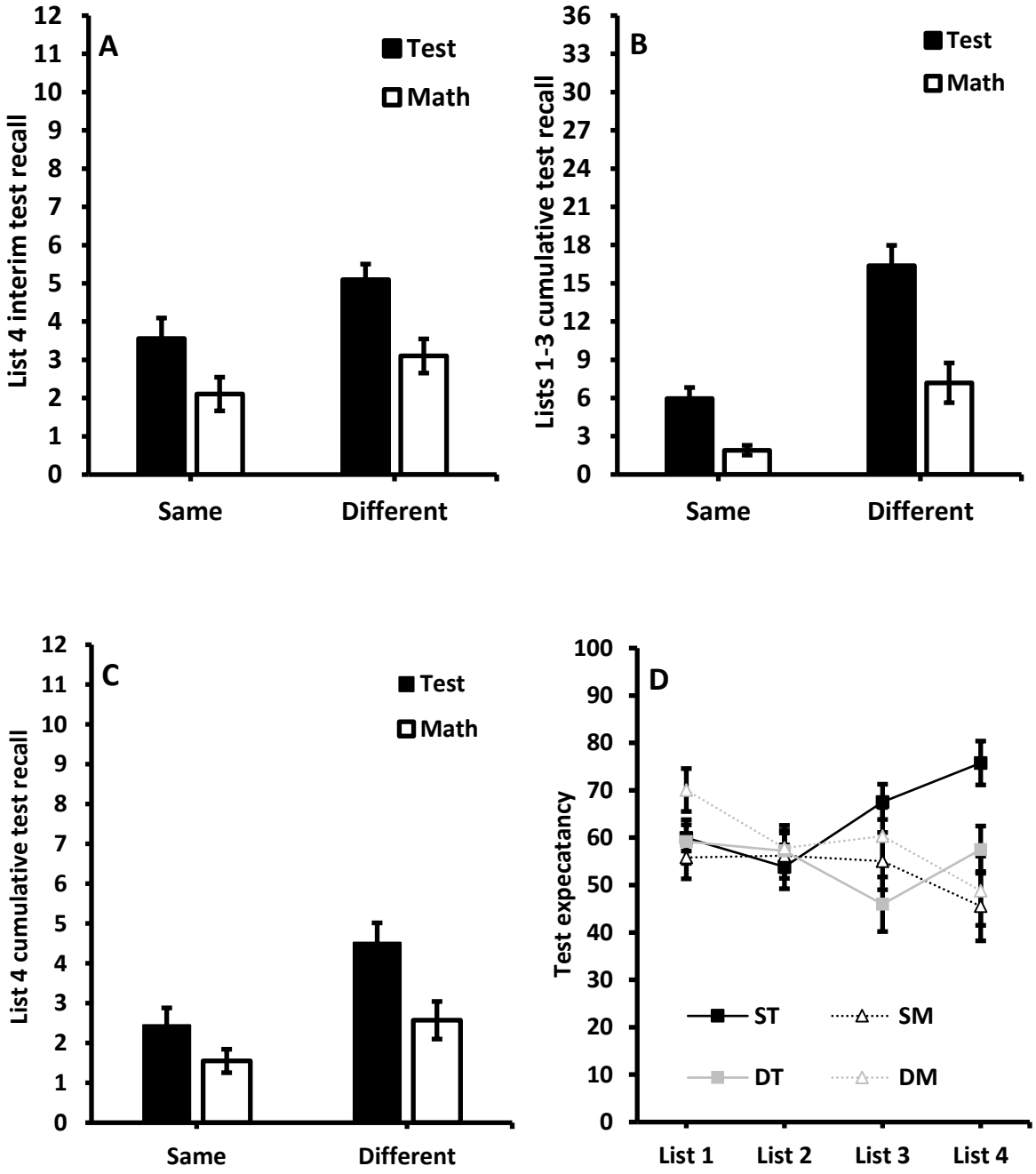

Figure 5.2: Experiment 7. Panel A: List 4 interim test recall; Panel B: Cumulative test recall of List 1-3 items; Panel C: Cumulative test recall of List 4 items; Panel D: Test expectancy ratings. ST = Same-Test; SM = Same-Math; DT = Different-Test; DM = Different-Math. Error bars represent $\pm$ 1 standard error.

information from a *different* domain, and that the forward testing effect is to some degree transferable.

An independent-samples *t* test revealed that the Same-Test group (M = 2.05, SD =1.61) suffered less from PI (i.e., incorrectly recalling another face's name from Lists 1-3) than the Same-Math group (M = 4.20, SD = 1.90), difference = -2.15 names, 95% CI = [-3.28, -1.02], Cohen's d = -1.22, indicating that interim testing prevents the build-up of PI. This result suggests that the classic testing effect (i.e., the difference in the List 4 interim test recall between the Same-Test and Same-Math groups) can partially be attributed to release from PI.

The amounts of PI in the Same-Test and Same-Math groups were significantly greater than 0, $t(19)$s > 5.71, ps < .001, Cohen's ds > 1.27. In contrast, neither the Different-Test nor the Different-Math groups experienced any PI. These results imply that the above finding, that a switch of material enhances recall of new information overall (i.e., the Different-Test and Different-Math groups outperformed the Same-Test and Same-Math groups in the List 4 interim test), might result from release from PI. Numerous previous studies have established that a switch of material type can enhance recall by reducing PI (e.g., Ellis & Montague, 1973; Emery et al., 2008; Lustig et al., 2001; Nunes & Weinstein, 2012). But more importantly, they also provide a strong challenge to the release from PI theory: PI was equivalent in the Different-Test and Different-Math groups, yet the former outperformed the latter in the critical List 4 test. Theoretical implications will be discussed more fully in the Discussion.

The four groups experienced roughly the same number of current list intrusions (i.e., incorrectly recalling another face's name from List 4): Same-Test: M = 2.15, SD = 1.73; Same-Math: M = 2.50, SD = 2.31; Different-Test: M = 2.62, SD = 2.50; Different-Math: M = 2.39, SD = 2.14. An ANOVA, with Material and Interim task as between-subjects variables,

revealed no main effect of Material, $F(1,78) = 1.65$, $p = .20$, $\eta_p^2 = .02$, no main effect of Interim task, $F(1,78) = 1.08$, $p = .30$, $\eta_p^2 = .01$, and no interaction, $F(1,78) = .11$, $p = .75$, $\eta_p^2 = .001$. The fact that the Same-Math group suffered more from PI than the Same-Test group, but these two groups experienced the same frequency of current list intrusions, replicates previous findings (Weinstein et al., 2011). It indicates that the Same-Test group's memory search set in the List 4 interim test was smaller and provides some support for the release from PI theory.

Cumulative test recall

Figure 5.2B depicts cumulative test recall across Lists 1-3. An ANOVA, with Interim task and Material as between-subjects variables, revealed a main effect of Material, $F(1,78) = 42.33$, $p < .001$, $\eta_p^2 = .35$, again indicating that Swahili-English pairs were easier to remember than face-name pairs. There was a main effect of Interim task, $F(1,78) = 30.02$, $p < .001$, $\eta_p^2 = .28$, confirming that interim testing enhances learning and retention more effectively than no interim testing (solving math problems). This might be caused by three possible reasons: (1) additional exposure to the recalled items (i.e., the Same-Test and Different-Test groups reviewed the recalled items in the interim tests); (2) the backward testing effect (i.e., testing of studied information enhances retention of studied/tested information compared to math problem solving); (3) the forward testing effect (i.e., prior interim tests enhance learning of Lists 2 and 3 compared to solving math problems). There was a significant interaction between Interim task and Material, $F(1,78) = 4.53$, $p = .04$, $\eta_p^2 = .06$, indicating that interim testing enhances retention of Swahili-English pairs somewhat more effectively than it does for face-name pairs.

Figure 5.2C depicts cumulative test recall on the List 4 items. An ANOVA with Interim task and Material as between-subjects variables revealed a main effect of Material,

$F(1,78) = 13.19$, $p = .001$, $\eta_p^2 = .15$, a main effect of Interim task, $F(1,78) = 11.20$, $p = .001$, $\eta_p^2 = .13$, but no interaction, $F(1,78) = 1.52$, $p = .22$, $\eta_p^2 = .02$. Although the interaction was nonsignificant, it showed the same pattern as recall of Lists 1-3.

Test expectancy ratings

Figure 5.2D depicts all four groups' test expectancy ratings across lists. The Same-Test and Different-Test groups gradually increased their test expectancy while the Same-Math and Different-Math groups gradually decreased their test expectancy across lists. Because the test expectancy ratings were noisy, the data were collapsed across groups to increase the power to observe possible effects: the Same-Test and Different-Test groups were collapsed as a Test group; the Same-Math and Different-Math groups were collapsed as a Math group. A mixed ANOVA, with Group (Test/Math) as a between-subjects variable and List (1-4) as a within-subjects variable, showed no main effect of Group, $F(1, 80) = 0.91$, $p = .34$, $\eta_p^2 = .01$, and no main effect of List, $F(1, 80) = 1.16$, $p = .29$, $\eta_p^2 = .01$. However, there was a significant linear interaction between Group and List, $F(1, 80) = 8.31$, $p = .005$, $\eta_p^2 = .09$. Follow-up repeated-measures ANOVAs with List as a within-subjects variable showed that the Test group linearly increased their test expectancy across lists but the linear trend did not reach significance, $F(1, 40) = 1.78$, $p = .19$, $\eta_p^2 = .04$, while the Math group linearly decreased their expectancy, $F(1, 40) = 7.22$, $p = .01$, $\eta_p^2 = .15$. The Test group reported higher test expectancy on List 4 than the Math group, difference = 19.17, 95% CI = [7.36, 30.98], Cohen's $d = 0.71$, but there was no significant difference on any of Lists 1-3, $t(80)$s < 0.82, ps > .42, Cohen's ds < 0.18. These results conceptually replicate Weinstein et al.'s (2014) findings and provide some support for the key process assumed to be critical according to test expectancy theory. The relationship between test expectancy and interim test recall will be discussed below.

Summary

The Same-Test group outperformed the Same-Math group in the List 4 interim test, replicating the classic forward testing effect. More interestingly, the Different-Test group also outperformed the Different-Math group in the List 4 interim test, revealing a degree of transferability of the forward testing effect. Test expectancy increased across lists in the Test groups but decreased in the Math groups, replicating Weinstein et al.'s (2014) test expectancy findings.

Experiment 8

The main aim of Experiment 8 was to extend Experiment 7's findings to different materials. In Experiment 8, the Same-Test, Same-Math, and Different-Math groups were omitted. A Different-Restudy group, which restudied material from the preceding block and was tested on the final (target) block, was added, which enabled this experiment to explore the transferability of the forward testing effect by comparing interim testing with restudying.

Experiment 7 illustrated that the forward testing effect is transferable when no corrective feedback was offered in any interim test. Unlike Experiment 7, in Experiment 8 corrective feedback was offered in all interim tests. There are both a theoretical and a practical rationale for this change. Providing corrective feedback equates the Interim Test and Interim Restudy groups in all respects (including re-exposure to the correct responses) except the critical one, namely whether interim tests are administered or not prior to the target (final) block. Giving corrective feedback, therefore, avoids possible influences from other non-targeted factors/variables (e.g., re-exposure to the correct responses) on the forward testing effect. The practical reason is that it would be unusual (and unpopular) to administer a test or quiz in a classroom or other learning environment without providing feedback. Hence, Experiment 8 employed a more naturalistic procedure.

In Experiment 7, both the Swahili-English and face-name pairs were paired-associates and the test format was cued-recall in all interim tests. Cho et al. (2016) suggested that interim tests may encourage people to adopt more effective encoding and retrieval strategies in the subsequent learning and recall phases because they provide information about the test format. Therefore, the transferability of the effect in Experiment 7 might result from encoding and retrieval strategy changes that support performance in the final target list. In Experiment 8, material type was changed from block to block: Block 1: object pictures; Block 2: text; Block 3: face-profession pairs. In addition, interim test format was also changed from block to block: Block 1: recognition; Block 2: fill-in-the-blank; Block 3: cued recall. These changes enabled this experiment to explore the transferability of the forward testing effect when material types and test formats are changed from block to block, thus minimizing any beneficial contribution from the strategy transfer mechanism proposed by Cho et al. (2016).

Test expectancy ratings in Experiment 7 were relatively noisy, which might have arisen from the fact that the rating scale (0-100) was too granular. In Experiment 8, the rating scale was narrowed (1-7). In Experiment 7, participants were asked to report how likely they thought it was that the computer would give them an interim test on the subsequent list, which might act as a test warning, reminding participants that they might be tested and encouraging them to exert more encoding effort. In Experiment 8, participants were instead asked to type in a number (1-7) to indicate what they thought the subsequent task would be: testing or restudying. They were informed: 1 = "*I am sure that the computer will offer me a restudy opportunity*"; 4 = "*I have no idea*"; 7 = "*I am sure that the computer will give me a test*".

Method

Participants

In Experiment 7, the effect size of the transferability of the forward testing effect was 0.84 (Cohen' *d*). The calculated sample size to observe a significant (α = .05; power = .90) forward testing effect in Experiment 8 was 29 participants in each group. Sixty-six undergraduates, mean age = 19.58 (*SD* = 0.96) years including 64 females, were recruited from Fuqing Branch of Fujian Normal University.[7] All participants' first language was Chinese and they completed this experiment for course credit. They were randomly allocated to two groups, with 32 in the Different-Test group and 34 in the Different-Restudy group. In this experiment, all text materials and instructions were in Mandarin.

Materials

Four hundred and fifty object pictures were selected from a published database developed by Brady, Konkle, Alvarez, and Oliva (2008) (available at http://cvcl.mit.edu/MM/stimuli.html). The test format for pictures was recognition, which was relatively easy, and people's memory capacity for image details is very substantial (Brady et al., 2008). To prevent a ceiling effect in the recognition test, the number of pictures was set to 450 in total. These pictures were randomly separated into three sets: the first set was used in Block 1's study phase; the first and second sets were used in Block 1's interim test and restudy phases; the first and third sets were used in Block 1's cumulative test phase. The order of these three sets was counterbalanced across participants using a Latin square design.

A science text concerning graphene was employed in Block 2. The text consisted of three paragraphs, each comprising ten sentences, and each sentence was roughly the same

[7] Experiments 7-9 were conducted as an international collaboration project. Therefore, the data in Experiment 8 were collected in China and those in Experiments 7 and 9 in the UK.

length. The first paragraph depicted the properties of graphene, the second paragraph concerned its uses, and the third paragraph was mainly about the research history of graphene.

Thirty Chinese male faces were selected from the CAS-PEAL face database developed by Gao et al. (2008) (available at http://www.jdl.ac.cn/peal/). Thirty common professions were selected from the *Dictionary of Occupations in China*. The faces and professions were randomly paired and the face-profession assignment was consistent across participants. These 30 face-profession pairs were used in Block 3.

Design and procedure

The experiment involved a between-subjects design (Interim task: test/restudy). Participants were instructed to study 150 pictures, a science text, and 30 face-profession pairs. All participants were warned of the cumulative test at the beginning of the experiment. They were also informed that the computer would randomly decide the following task after studying each block and solving math problems for 60 sec: a short test or restudying the prior block. In fact, the Different-Test group undertook interim tests on all three blocks, whereas the Different-Restudy group restudied Blocks 1 and 2 and undertook an interim test on Block 3 (see Figure 5.3).

Prior to each block's study phase, participants were instructed to type in a number (1-7) to indicate their expectancy of the next task. In Block 1's study phase 150 pictures were presented one by one, for 2 sec each, in a random order. Next, both groups solved math problems for 60 sec. Then the Different-Test group took an interim test, in which 300 (150 studied and 150 new) pictures were randomly presented one by one and participants' task was to judge whether the presented picture was old (studied) or new. Corrective feedback ("old" or "new") was shown for 1 sec following each response. In contrast, the Different-Restudy

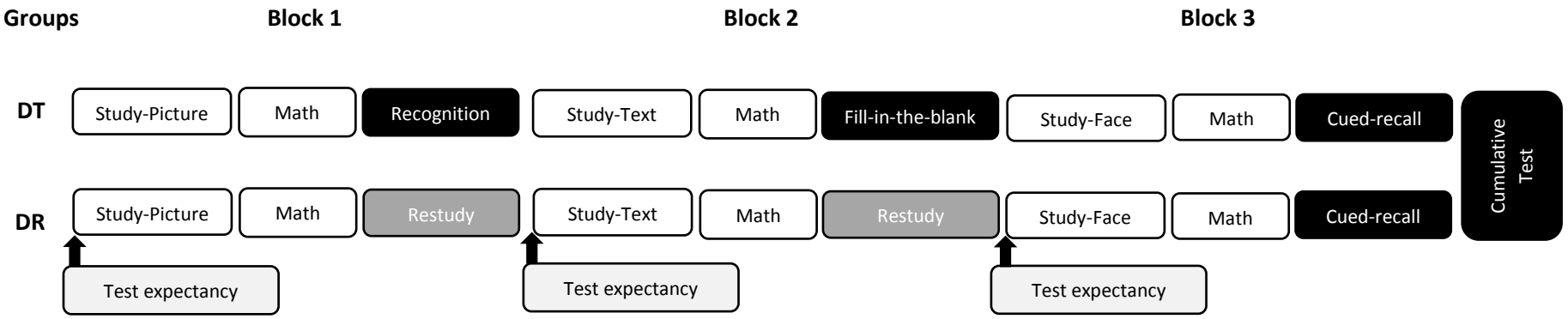

Figure 5.3: Experiment 8. The Different-Test (DT) and Different-Restudy (DR) groups studied different types of material across three blocks: Block 1: object pictures; Block 2: text; Block 3: face-profession pairs. Prior to studying each block, both groups reported their test expectancy. The Different-Test group took interim tests on all three blocks whereas the Different-Restudy group restudied Block 1 and 2 items and took an interim test on Block 3. The test formats changed from block to block: Block 1: recognition; Block 2: fill-in-the-blank; Block 3: cued recall. Both groups took a cumulative test.

group studied the same 300 pictures. These pictures were shown one by one, for 2 sec each, in a random order, with "old" (for studied pictures) or "new" (for new pictures) presented below.

In Block 2's study phase, the entire text was shown on screen for 300 sec for participants to study. After solving math problems for 60 sec, the Different-Test group took a fill-in-the-blank test. Thirty sentences with target items omitted (e.g., *Graphene is about ____ times stronger than the strongest steel*) were presented one by one in a fixed sequence (i.e., the same sequence as they appeared in the text). The target item in each sentence was a digit number or a two-character Chinese word. Participants were asked to type their answers into a blank box. Following each response, the correct answer (e.g., *200*) was presented for 3 sec as corrective feedback. The Different-Restudy group restudied the entire text. The 30 sentences were presented one by one, for 10 sec each, in the same sequence as they appeared in the text. The target item in each sentence was underlined and in red.

In Block 3's study phase, the 30 face-profession pairs were presented one by one, in a random order, for 10 sec each. After solving math problems for 60 sec, both the Different-Test and Different-Restudy groups undertook a cued recall test, in which the 30 faces were presented one by one in a new random order. Participants were instructed to recall their corresponding professions. Following each response, the correct profession was presented for 3 sec as corrective feedback.

Following the completion of Block 3, both groups undertook a cumulative test. There was no feedback in this test. Participants took a recognition test first, in which 300 (150 studied in Block 1's study phase and 150 completely new) pictures were shown one by one in a random order and participants were asked to make old/new judgments. Next, they took a fill-in-the-blank test on the text. The sentences without target items were presented one by

one in the same sequence as they appeared in the text and participants were asked to recall the targets. Finally, they completed a cued recall test on all 30 face-profession pairs. The faces were presented one by one in a new random order and participants were asked to recall the associated professions. The experiment lasted about 55 min.

Results

Interim test performance on Blocks 1 and 2

In the Block 1 interim test, the Different-Test group's mean hit (i.e., judging studied pictures as old) rate was 72.6% (*SD* = 12.46) and mean false alarm (i.e., mistakenly judging new pictures as old) rate was 22.8% (*SD* = 14.37). Discrimination (i.e., discriminating studied from new pictures) was significantly greater than 0, d' = 1.50, 95% CI = [1.20, 1.80]. In the Block 2 interim test, participants correctly recalled 13.28 (*SD* = 10.01) out of 30 target items.

Block 3 interim test recall

Of critical interest is the difference in Block 3 interim test recall between groups. An independent-samples *t* test found that the Different-Test group (*M* = 8.50, *SD* = 6.54) correctly recalled about twice as many professions as the Different-Restudy group (*M* = 4.89, *SD* = 4.17), difference = 3.61 professions, 95% CI = [0.94, 6.30], Cohen's *d* = 0.66 (see Figure 5.4A), again revealing that the forward testing effect is robustly transferable.

Cumulative test performance

For Block 1 items in the cumulative test, as can be seen in Figure 5.4B, both groups were able to discriminate studied from new pictures: for the Different-Test group: d' = 2.08, 95% CI = [1.77, 2.39]; for the Different-Restudy group: d' = 1.51, 95% CI = [1.13, 1.89]. More importantly, discrimination was better in the Different-Test group than in the Different-Restudy group, difference in d' = 0.57, 95% CI = [0.09, 1.06], Cohen's *d* = 0.59. This result

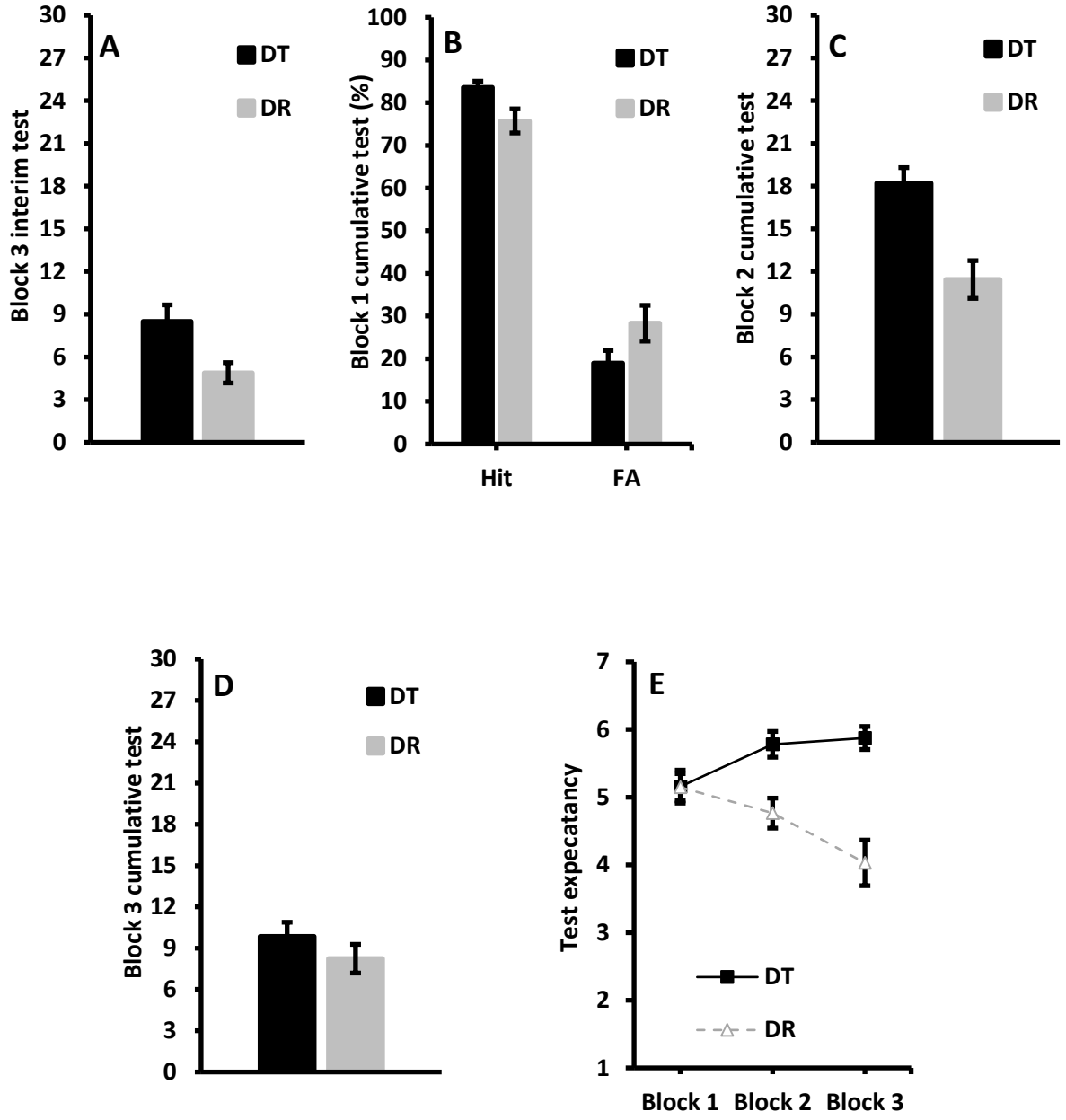

Figure 5.4: Experiment 8. Panel A: Block 3 interim test recall; Panel B: Hit and false alarm (FA) rates in the cumulative test for Block 1 items; Panel C: Cumulative test recall for Block 2 items; Panel D: Cumulative test recall for Block 3 items; Panel E: Test expectancy ratings. DT = Different-Test; DR = Different-Restudy. Error bars represent ± 1 standard error.

indicates a clear backward testing effect on recognition memory (Jacoby et al., 2010).

For Block 2 items in the cumulative test, the Different-Test group (M = 18.22, SD = 6.12) recalled substantially more target items than the Different-Restudy group (M = 11.44, SD = 7.75), difference in recall = 6.78 items, 95% CI = [3.33, 10.23], Cohen's d = 0.97 (see Figure 5.4C). This difference might result from both forward and backward testing effects: (1) the Block 1 interim test, compared to restudying Block 1 items, might have motivated the Different-Test group to commit more effort toward encoding Block 2 material; (2) the Block 2 interim test might have enhanced retention more efficiently than restudying Block 2.

For Block 3 items in the cumulative test, the Different-Test group (M = 9.84, SD = 5.89) recalled numerically (but not significantly) more professions than the Different-Restudy group (M = 8.24, SD = 6.09), difference in recall = 1.61 professions, 95% CI = [-1.34, 4.56], Cohen's d = 0.27 (see Figure 5.4D), the same qualitative pattern as observed in the Block 3 interim test. Previous studies showed that testing enhances subsequent encoding of corrective feedback (Butler & Roediger, 2008; Potts & Shanks, 2014). As can be seen in Figures 5.4A and 5.4D, the Different-Restudy group benefited much more from the Block 3 interim test than the Different-Test group. Nonetheless, the interim test with corrective feedback (Block 3 interim test) was insufficient to completely overcome the forward testing effect, emphasizing the robustness of the effect (Szpunar et al., 2013).

Test expectancy ratings

Mean test expectancy is shown in Figure 5.4E. A mixed ANOVA, with Block (1-3) as a within-subjects variable and Interim task as a between-subjects variable, showed that test expectancy was higher in the Different-Test group than the Different-Restudy group, $F(1, 64) = 12.89$, $p = .001$, $\eta_p^2 = .17$, but there was no main effect of Block, $F(1, 64) = 0.81$, $p = .37$, $\eta_p^2 = .01$. Importantly, there was a significant linear interaction between Block and Interim

task, $F(1, 64) = 17.10$, $p < .001$, $\eta_p^2 = .21$. Test expectancy linearly increased across blocks in the Different-Test group, $F(1, 31) = 8.27$, $p = .002$, $\eta_p^2 = .28$, but linearly decreased in the Different-Restudy group, $F(1, 33) = 8.47$, $p = .006$, $\eta_p^2 = .20$. Independent-samples *t* tests showed no difference in test expectancy between the groups in Block 1, difference = 0.01, 95% CI = [-0.63, 0.64], Cohen's $d = 0.006$, but the difference was significant in Block 2, difference = 1.02, 95% CI = [0.42, 1.62], Cohen's $d = 0.84$, and Block 3, difference = 1.85, 95% CI = [1.04, 2.65], Cohen's $d = 1.44$. The relationship between test expectancy and interim test recall will be discussed more fully below.

Summary

Experiment 8 again revealed that the forward testing effect is transferable. Going beyond Experiment 7, Experiment 8 demonstrates substantial transfer even when material types and test formats are changed from block to block. Test expectancy ratings again conceptually replicated Weinstein et al.'s (2014) test expectancy findings.

Experiment 9

Experiments 7 and 8 demonstrated that the forward testing effect is transferable across different domains of relatively low-level learning (i.e., item learning). The main aim of Experiment 9 is to explore whether the effect transfers from low- to high-level learning (e.g., inductive learning; for a detailed discussion about the differences between low- and high-level learning, see Chapter 4). The second aim of Experiment 9 is to test the transferability of the forward testing effect using more educationally-realistic materials and in a simulated classroom setting.

In order to probe in detail the correlation across participants between final list/block test expectancy and interim test recall (see below), a large sample size is required. Therefore,

the third aim of Experiment 9 is to increase the sample size to further examine the role of test expectancy.

As noted previously, a retrieval-effort mechanism may contribute to the forward testing effect. However, this theory has not been directly examined yet, therefore the fourth aim of Experiment 9 is to test this theory. It hypothesizes that retrieval failures in prior interim tests motivate individuals to increase their retrieval effort. To test this theory, participants' response times (RTs) in the test stage of the final (target) block were measured. RTs were taken as an index of retrieval effort. According to the retrieval-effort theory, the Different-Test group should exert more effort (indexed by longer RTs) to answer the questions than the Different-Restudy group in the target block test.

Experiment 8 demonstrated that interim testing with corrective feedback enhances subsequent learning more effectively than restudying. In Experiment 9, corrective feedback in the interim tests was omitted, which allowed this experiment to directly compare the effect of interim testing with that of restudying on subsequent learning and retrieval of new information, removing any influences from additional learning via corrective feedback.

Unlike in Experiments 7 and 8, the final test was omitted in Experiment 9. The main interest of Experiment 9 is the Block 4 interim test performance, and there was a class time limit for the experiment. Numerous previous studies have documented that testing of studied information enhances its retention by comparison with restudying (for a review, see Roediger & Karpicke, 2006a), and the Same-Test and Different-Test groups consistently outperformed the Same-Math, Different-Math, and Different-Restudy groups in the final tests in Experiments 7 and 8 – the same patterns repeatedly documented in many previous forward testing effect studies (e.g., Jing et al., 2016; Szpunar et al., 2014; Szpunar et al., 2013; Szpunar et al., 2008; Weinstein et al., 2014; Experiments 1-6 in this thesis).

Method

Participants

One hundred and thirty-eight UCL first-year psychology undergraduate students were recruited from an Experimental Psychology class. They participated as a course requirement and the sample size was determined by the class size. Six participants' data were not recorded because of computer failure, leaving a final sample of 132 participants (mean age = 18.89 years, *SD* = 1.40; 111 females; 86 participants' first language was English). They were randomly separated into two groups, with 64 in the Different-Test group and 68 in the Different-Restudy group. According to the effect size in Experiment 8 (Cohen's $d = 0.66$), the power to observe a significant ($\alpha = .05$) forward testing effect in Experiment 9 is about 0.97.

Materials

The principal stimuli in Blocks 1-3 were 30 statements about famous artists (available at http://www.oil-painting-techniques.com/history-of-oil-painting.html). Each statement was a short sentence, describing an artist's contributions to art (e.g., *Veronese introduced a greater realism and sumptuous, decorative color*). These statements were randomly divided into three sets, with 10 statements in each set, and these three sets were assigned to Blocks 1-3.

The stimuli in Block 4 were 80 paintings comprising 10 from each of eight artists (e.g., Philip Juras, Ryan Lewis). The paintings, which were relatively unknown to students, were taken from Kornell and Bjork (2008a). Forty-eight paintings, consisting of six paintings from each of the eight artists, were used in Block 4's study phase and these were separated into six sets, each set consisting of one painting by each artist. The other 32 paintings were used in the Block 4 test.

Design and procedure

The experiment employed a between-subjects design (Interim task: test/restudy). Experiment 9 was conducted in a laboratory classroom. Participants were group-tested (up to 50 participants in each group) on personal computers. At the end of the experiment, participants were instructed not to discuss the experiment with their classmates.

Participants were instructed to imagine themselves as an art student taking an art class involving four blocks of learning. They were encouraged to remember as much information as they could. They were also informed that, after studying each block, the computer would decide whether to offer them a restudy opportunity or give them a short test. In fact, the Different-Test group took a test on every block whereas the Different-Restudy group restudied Blocks 1-3 and took a test on Block 4. Performance on the Block 4 test is the main dependent measure.

Figure 5.5 schematically illustrates the design. Before studying each block, participants reported their test expectancy on a slider ranging from 1 ("*I am sure the computer will offer me a restudy opportunity*") to 9 ("*I am sure the computer will test me*"). In Block 1's study phase, the 10 statements were presented one by one in a random order, for 20 sec each. Following the study phase, the Different-Test group took a fill-in-the-blank test on these statements. The 10 statements were presented one by one, in a new random order, with a word or phrase missing in each statement (e.g., *Veronese introduced a greater* _____ *and sumptuous, decorative color*). Participants were asked to recall and type their answers into a blank box. They had up to 20 sec to answer each question and they were allowed to leave the question empty if they were unable to recall the correct answer. By contrast, the Different-Restudy group restudied these 10 statements one by one in a new random order, for 20 sec each. In Blocks 2 and 3 participants performed the same task as in Block 1, except that

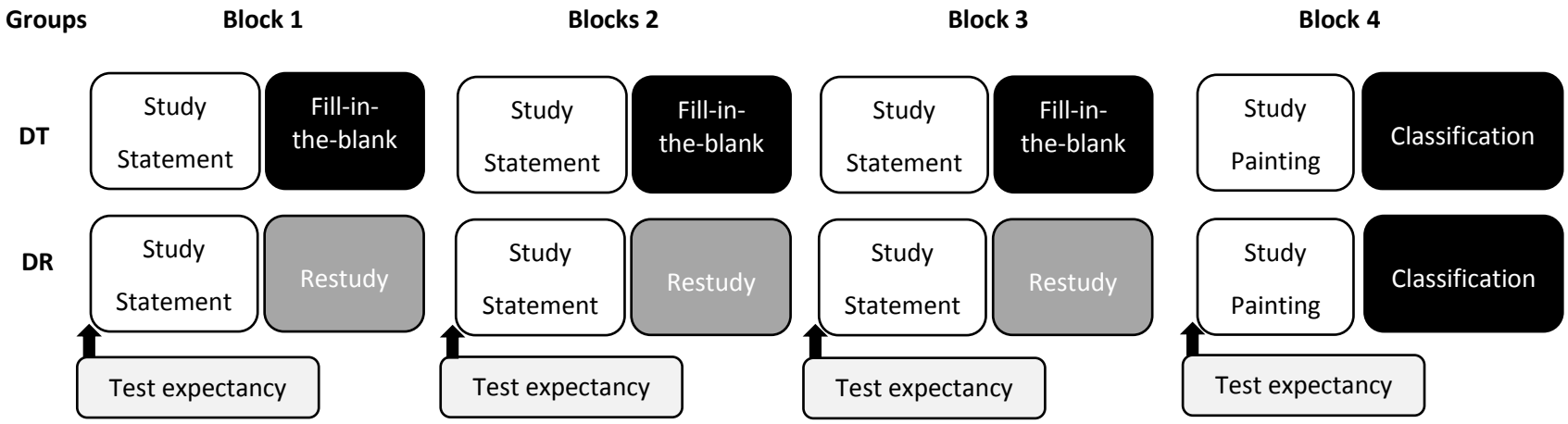

Figure 5.5: Experiment 9. The Different-Test (DT) and Different-Restudy (DR) groups studied three blocks of statements followed by a block of paintings. Prior to studying each block, both groups reported their test expectancy. The Different-Test group took tests on all four blocks whereas the Different-Restudy group restudied Block 1-3 items and took a test on Block 4.

they studied 10 new statements.

In the Block 4 study phase, the 48 paintings were presented one by one, for 5 sec each, with the artist's name presented below. The paintings were presented in a spaced arrangement, following Kornell and Bjork (2008a). All the pictures from one set (consisting of 1 painting by each of the 8 artists) were shown, then those from the next set, and so on, in an order that was fixed for all participants. Following the study phase, both groups were tested on their ability to attribute new paintings to the artists. The 32 test paintings were presented one by one in a random order, with the eight artists' names presented below each painting. Participants were asked to choose who the corresponding artist was for each painting and they had up to 20 sec to respond. The Experiment lasted about 30 min.

Results

Scoring

In the Block 1-3 interim tests, 2 points were assigned to correct answers and 1 point to partially correct answers. An assessor scored the Block 1-3 test responses for the Different-Test group. Only one assessor scored Block 1-3 test responses for the Different-Test group for the following reasons: (1) the Different-Restudy group did not take tests on Blocks 1-3; (2) no comparison was made between groups on their Block 1-3 test performance; (3) Block 1-3 test performance is not a key outcome measure. The main interest is the Block 4 test performance, which was automatically scored by the computer for both the Different-Test and Different-Restudy groups.

Block 1-3 test recall

In the Block 1-3 interim tests, the Different-Test group's scores were 4.55 (*SD* = 2.94), 4.39 (*SD* = 2.91), and 4.67 (*SD* = 2.95), respectively out of 20.

Block 4 test performance

Of critical interest is accuracy on the Block 4 test. The Different-Test group (M = 23.92, SD = 4.13) correctly classified more paintings than the Different-Restudy group (M = 19.49, SD = 7.95), difference = 4.44 paintings, 95% CI = [2.23, 6.64], Cohen's d = 0.66 (see Figure 5.6A), revealing that the forward testing effect is robustly transferable from verbal fact to visual concept learning. A mixed ANOVA, with Interim task and Language (first language: English or other) as between-subjects variables, was conducted to explore whether Language moderated the transferability of the forward testing effect. This yielded a main effect of Interim task, $F(1, 128) = 16.12$, $p < .001$, $\eta_p^2 = .11$, but there was no main effect of Language, $F(1, 128) = 1.06$, $p = .30$, $\eta_p^2 = .007$, and no interaction between Interim Task and Language, $F(1, 128) = 0.03$, $p = .87$, $\eta_p^2 < .001$. Hence language did not significantly moderate the transfer.

Retrieval effort in the Block 4 test

Mean RT in the Block 4 test was calculated for each participant (see Figure 5.6B). An independent-samples t test showed that the Different-Test group took longer (M = 3411 ms, SD = 929) to classify pictures than the Different-Restudy group (M = 3108 ms, SD = 568), difference = 303 ms, 95% CI = [40, 566], Cohen's d = 0.40, consistent with the retrieval-effort theory.[8]

Further analyses were conducted to explore whether classification accuracy (correct/incorrect) moderated the effect of prior interim tests (i.e., Block 1-3 tests) on

[8] Median RTs showed the same pattern: The Different-Test group spent longer (M = 2839 ms, SD = 713) than the Different-Restudy group (M = 2610 ms, SD = 543), difference = 228 ms, 95% CI = [11, 446], Cohen's d = 0.36.

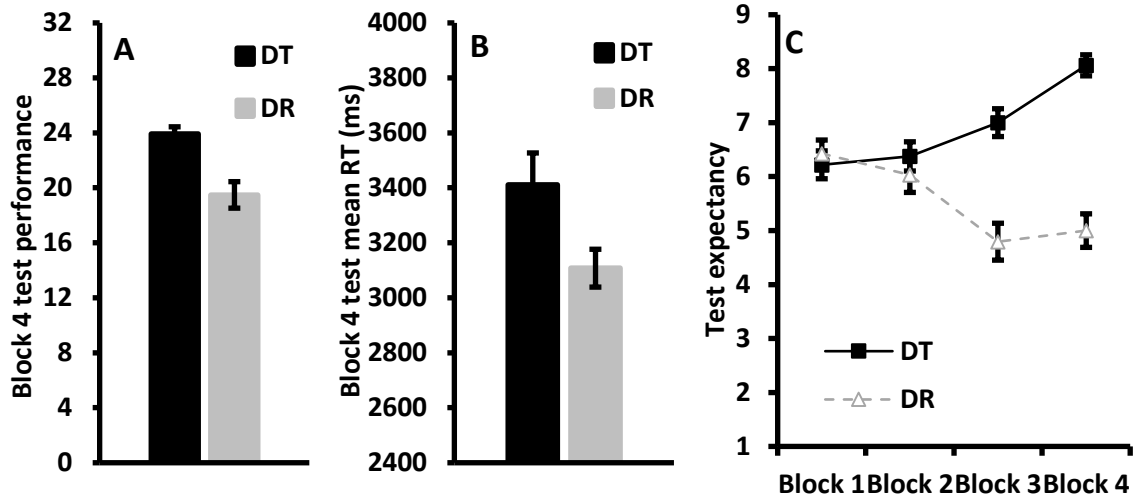

Figure 5.6: Experiment 9. Panel A: Block 4 test performance; Panel B: Mean RTs in the Block 4 test; Panel C: Test expectancy ratings. DT = Different-Test; DR = Different-Restudy. Error bars represent ± 1 standard error.

subsequent retrieval effort (i.e., RTs in the Block 4 test). Mean RTs for correctly and incorrectly classified paintings were calculated for each participant, and a mixed ANOVA with Classification accuracy (correct/incorrect) as a within-subjects variable and Interim task as a between-subjects variable was conducted. The results showed a main effect of Interim task, $F(1, 130) = 3.93$, $p = .049$, $\eta_p^2 = .03$, again indicating that the Different-Test group exerted more retrieval effort than the Different-Restudy group. There was also a main effect of Classification accuracy, $F(1, 130) = 69.75$, $p < .001$, $\eta_p^2 = .35$: participants responded faster to correctly classified paintings than to incorrectly classified ones.

There was no significant interaction between Interim task and Classification accuracy, $F(1, 130) = 0.93$, $p = .34$, $\eta_p^2 = .005$, indicating that classification accuracy did not significantly moderate the effect of prior interim tests on subsequent retrieval effort. This conclusion should be treated with caution however because there were more correctly than

incorrectly classified paintings (i.e., unequal numbers of correctly and incorrectly classified paintings).

Although participants responded faster to correctly classified paintings and the Different-Test group classified more paintings correctly than the Different-Restudy group, the Different-Test group still spent more time on classification. The difference in classification accuracy did not eliminate the difference in RTs between the groups, revealing the robustness of the difference in retrieval effort between groups and supporting the retrieval-effort theory.

Test expectancy ratings

Mean test expectancy ratings, shown in Figure 5.6C, evolved differently for the Different-Test and Different-Restudy groups. As anticipated, participants in the Different-Test group developed an increasing expectation of being tested while those in the Different-Restudy group showed an increasing expectation of a restudy opportunity. A mixed ANOVA, with Block (1-4) as a within-subjects variable and Interim task as a between-subjects variable, found a main effect of Interim task, $F(1, 130) = 34.37$, $p < .001$, $\eta_p^2 = .25$. There was no main effect of Block, $F(1, 130) = 0.81$, $p = .67$, $\eta_p^2 = .001$, but the interaction between Block and Interim task was significant, $F(1, 130) = 59.85$, $p < .001$, $\eta_p^2 = .32$. Test expectancy increased linearly across blocks in the Different-Test group, $F(1, 63) = 43.31$, $p < .001$, $\eta_p^2 = .41$, but decreased linearly in the Different-Restudy group, $F(1, 67) = 22.83$, $p < .001$, $\eta_p^2 = .25$.[9] Independent-samples *t* tests showed no significant difference in test expectancy between groups in Blocks 1, difference = -0.36, 95% CI = [-0.92, 0.51], Cohen's $d = -0.10$, and 2, difference = 0.35, 95% CI = [-0.49, 1.18], Cohen's $d = 0.14$. However, there

[9] The cubic trend of the Same-Restudy group's test expectancy across blocks was also significant, $F(1, 67) = 4.15$, $p = .046$, $\eta_p^2 = .06$.

were significant differences in Blocks 3, difference = 2.21, 95% CI = [1.35, 3.01], Cohen's *d* = 0.89, and 4, difference = 3.06, 95% CI = [2.33, 3.80], Cohen's *d* = 0.89. The relationship between test expectancy and test performance will be discussed below.

Summary

The forward testing effect is transferable from low- (verbal text) to high- (visual concept) level learning. Prior interim tests motivated participants to exert greater effort toward retrieving new information, consistent with the retrieval-effort theory. Test expectancy ratings again conceptually replicated Weinstein et al.'s (2014) findings.

The relationship between test expectancy and test performance

All three experiments consistently showed that the test groups (i.e., the Same-Test group in Experiment 7 and the Different-Test groups in Experiments 7-9) reported higher test expectancy than the control groups (i.e., the Same-Math and the Different-Math groups in Experiment 7 and the Different-Restudy groups in Experiments 8 and 9) on the target list/block. To determine directly whether test expectancy contributes to the forward testing effect, correlation analyses on the relationship between test expectancy and interim test performance were conducted.

The List 4 test expectancy ratings and test recall data were collapsed across the Same-Test and Same-Math groups in Experiment 7. These were significantly correlated, $r_{(40)} = .35$, $p = .03$, revealing an association between test expectancy and recall, consistent with the idea that test expectancy contributes to the same-materials forward testing effect.

To explore whether test expectancy contributes to the transferability of the forward testing effect, the List 4 expectancy ratings and test recall data were collapsed across the Different-Test and Different-Math groups in Experiment 7. These were not significantly correlated, $r_{(40)} = .11$, $p = .47$. Similarly, the Block 3 expectancy ratings and recall data were

collapsed across the Different-Test and Different-Restudy groups in Experiment 8: $r_{(66)} = .18$, $p = .16$. Lastly, there was also no significant correlation between Block 4 expectancy ratings and test performance in Experiment 3, $r_{(132)} = .13$, $p = .13$. Figure 5.7 depicts the associations between test expectancy and recall in Experiments 7-9.

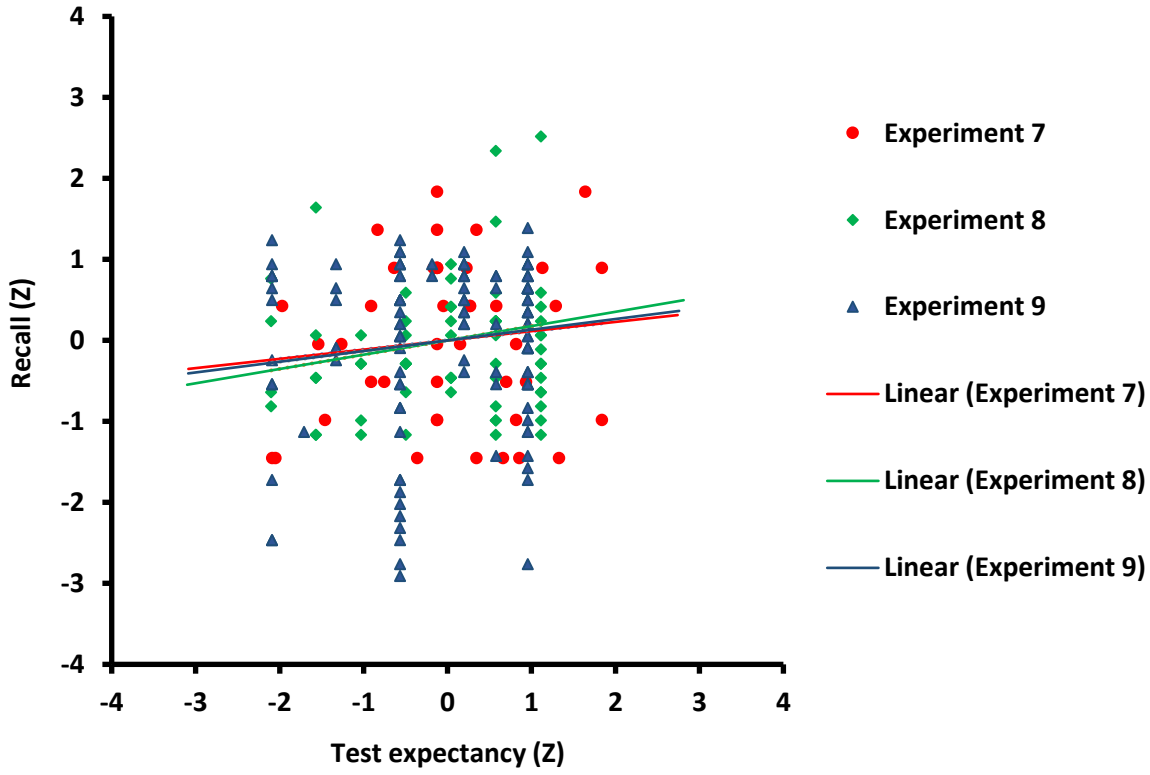

Figure 5.7: Scatter plot and linear trends between test expectancy and recall in Experiments 7-9 (Experiment 7's Same-Test and Same-Math groups were excluded). Given that the test expectancy rating scales were different, test expectancy ratings and recall data were transformed into *Z* scores in each experiment.

These results imply no significant correlation between test expectancy and test performance, challenging the proposal that test expectancy contributes to the transferability of the forward testing effect. However, it is possible that these non-significant correlations

are false negatives arising from inadequate sample sizes and low statistical power (Vadillo, Konstantinidis, & Shanks, 2016). Across all three experiments, the results consistently showed a positive (although non-significant) correlation between test expectancy ratings and test performance. Therefore, a meta-analysis was conducted to increase power.

In the meta-analysis, the Same-Test and Same-Math groups from Experiment 7 were excluded because the main aim of this meta-analysis is to explore whether test expectancy contributes to the transferability of the forward testing effect.[10] Using formulae explained by Borenstein, Hedges, Higgins, and Rothstein (2009), the *r* values were first transformed into Cohen's *d*s using the formula:

$$d = \frac{2r}{\sqrt{1 - r^2}}$$

The variances of the *r* values were calculated using the formula:

$$V_r = \frac{1 - r^2}{N}$$

where *N* represents the sample size. Next, the variances of Cohen's *d*s were calculated using the formula:

$$V_d = \frac{4V_r}{(1 - r^2)^3}$$

These Cohen's *d*s and V_ds were then inserted into the R *metafor* package and a random-effects meta-analysis was conducted. This revealed a significant albeit modest effect of test

[10] Including Experiment 7's Same-Test and Same-Math groups does not change the pattern of results.

expectancy on recall, Cohen's d = 0.28, 95% CI = [0.03, 0.54] (see Figure 5.8 for detailed results).[11]

Finally, this effect size (Cohen's d) was transformed back to r using the formula:

$$r = \sqrt{\frac{d^2}{d^2 + 4}}$$

This yielded an r value of .14, confirming a weak correlation between test expectancy

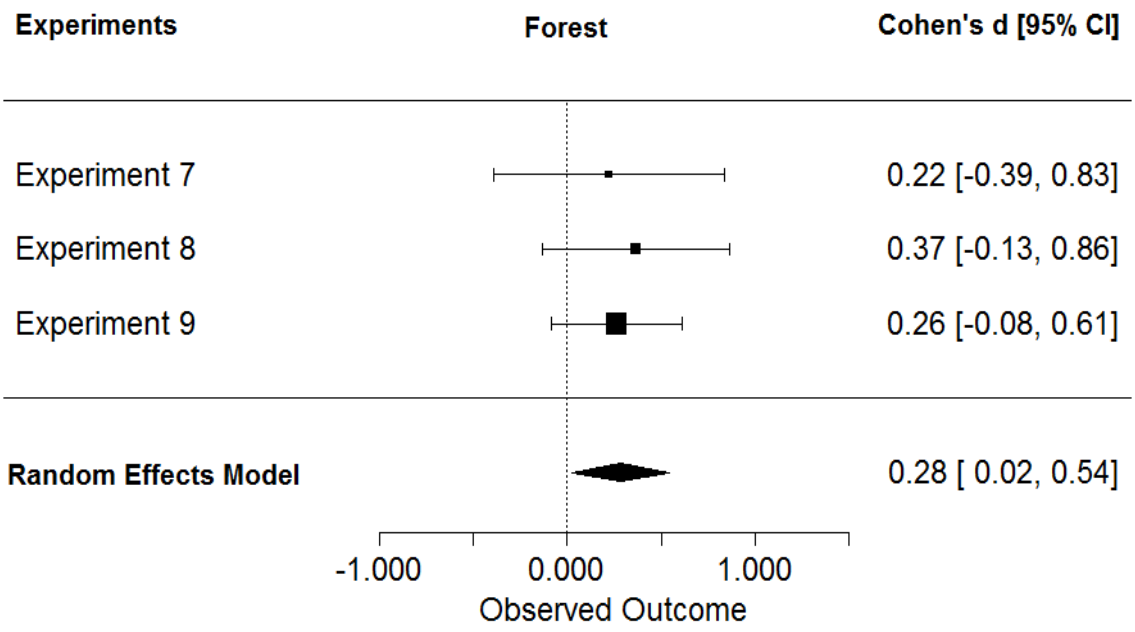

Figure 5.8: Forest plot of the meta-analysis of the effect of test expectancy on test performance. Error bars represent 95% CI.

[11] Given that Experiment 7's test expectancy ratings were relatively noisy, a new random effects meta-analysis was conducted, in which the Different-Test and Different-Math groups from Experiment 7 were excluded. The new meta-analysis also showed a significant effect of test expectancy on test performance, Cohen's d = 0.30, 95 % CI = [0.01, 0.58], p = .04. The transformed r value is .15.

and test performance.[12]

In summary, the significant correlation between test expectancy and recall performance in the Same groups (Same-Test and Same-Math) in Experiment 7 supports the test-expectancy theory as an account for the standard forward testing effect. The above meta-analysis reveals a significant, although small, effect of test expectancy on test performance when the material is changed, supporting the test-expectancy theory as an account for the transferability of the forward testing effect.

Discussion

Many previous studies have documented that testing of studied information from one domain enhances learning and retention of new information within the same domain (e.g., Szpunar et al., 2013; Szpunar et al., 2008). The current chapter goes beyond this to ask whether testing of studied information from one domain enhances learning and retention of new information from a different domain. In Experiment 7, the Same-Test group correctly recalled more names than the Same-Math group in the List 4 interim test, revealing a typical forward testing effect. The Same-Test group was less affected by PI in the List 4 interim test than the Same-Math group, while these two groups committed about the same number of current list intrusions. More novel was the finding that the Different-Test group recalled more names correctly in the List 4 interim test than the Different-Math group, revealing that the forward testing effect is transferable. Because of the switch of material types, neither the Different-Test nor Different-Math groups experienced any PI in the List 4 interim test.

[12] Besides meta-analysis, an alternative method is to apply Fisher's *Z* transformation and then calculate the correlation across all the data (Silver & Dunlap, 1987). This method again showed a significant correlation between test expectancy and test performance: $r = 0.14$, $p = .024$.

Experiment 8 again confirmed this transfer. More importantly, it revealed substantial transfer even when both the material type and test format are changed from block to block. Experiment 9 demonstrated that the forward testing effect is transferable from relatively low-level (fact learning) to high-level (visual concept) learning. In all three experiments, test expectancy increased across lists in the test groups (the Same-Test and Different-Test groups in Experiments 7-9) but decreased in the control groups (the Same-Math and Different-Math groups in Experiment 7 and the Different-Restudy groups in Experiments 8 and 9). These test expectancy ratings conceptually replicated Weinstein et al.'s (2014) findings. Furthermore, there were significant albeit modest correlations between test expectancy and test performance.

Theoretical implications

Turning to the theoretical interpretation of the results, several theories have difficulty explaining the transferability of the forward testing effect (i.e., the forward testing effects observed in Experiments 7-9's Different groups). As discussed above, the release-from-PI, activation-facilitation, and encoding-reset mechanisms should contribute little to transfer. Moreover, although the encoding/retrieval-strategy-change mechanism might have contributed to the transfer findings in Experiment 7 – because Swahili-English word pairs and face-name pairs were both paired-associates and the test formats were always cued-recall in all interim tests. The encoding and retrieval strategy theories have difficulty explaining the transfer findings in Experiments 8 and 9, in which material types and test formats were both switched, because there was little reason to expect that the strategies developed in prior learning and testing phases would be applicable to aid learning and retrieval in a different domain.

In contrast, the test-expectancy theory readily explains the transfer findings in Experiments 8 and 9. These experiments consistently showed that prior interim tests induced participants to expect an interim test on the next list/block, which might motivate them to exert more encoding effort. The positive correlation between test expectancy and test performance supplied additional evidence supporting the test-expectancy theory. The correlation between test expectancy and recall was modest, which could be interpreted as indicating that the role of test expectancy in the transfer of the forward testing effect is small. However, the weak correlations might result from measurement error. The test expectancy scores were composed of the true test expectancy plus error, but given that test expectancy was measured by a single question in the final list/block, the scores might be largely composed of error. Future research is encouraged to employ more reliable methods (e.g., measuring test expectancy repeatedly using different questions and calculating the mean across different questions).

Besides test-expectancy, retrieval-effort also appears to contribute to the transfer. As shown in Experiment 9, prior interim tests induced participants to spend longer retrieving the target information. Experiment 9 is the first to directly test the retrieval-effort theory. It must be noted that, although the results support the test-expectancy and retrieval-effort theories, the current chapter does not exclude other (e.g., encoding-reset, strategy-change) theories because these are not mutually exclusive.

There remain important questions about the transfer of the forward testing effect. For example, the time duration over which it operates is unknown. In the present chapter, as in all previous studies, the effects of interim tests have been evaluated at very short intervals. But if each list and test was separated by an interval of a day, for example, would transfer of the forward testing effect (and indeed the same-materials forward testing effect) still occur?

Another important question is whether the effect is modulated by test difficulty. Based on the failure-encoding-effort theory, more difficult tests (which induce more retrieval failures and motivate greater encoding effort toward subsequent encoding) may produce larger forward benefits.

Practical (educational) implications

Learners' study effort (e.g., attention) and learning effectiveness tend to decay across a study phase, and attenuated study effort leads to a decline in learning efficiency and impairs learning outcomes. How to sustain learners' study effort and learning effectiveness across a learning episode such as a class or lecture is a key concern for learners, educators, and psychologists. Experiment 7 confirmed that interim testing of studied information from one domain enhances encoding and retention of new information from the same domain, indicating that interim tests can be employed as a practical strategy to enhance the learning of new information while studying additional material of the same type.

In natural learning situations, to-be-studied content frequently varies, which highlights the importance of exploring the transferability of the forward testing effect. Experiment 7 demonstrated transfer; Experiment 8 revealed that the forward testing effect transfers even when material types and test formats are changed from block to block; and Experiment 9 demonstrated transfer from low- to high-level learning. This successful transfer, repeatedly observed in three experiments, suggests that interim tests can be employed to improve learning of new information while studying additional material of a different type. Overall, the current chapter suggests that interim testing can be profitably used to enhance learning and retention of new information from both the same and different domains.

Students frequently suffer from PI in educational settings. For example, in a history class, high school students need to remember the dates of different historical events. They may confuse a newly studied event's date with those of other studied events. People also frequently suffer from PI in daily life. For example, imagine that you are attending a party, in which you are about to meet a few new people and you need to commit their names to memory. You might confuse a new person's name with other persons' names (Weinstein et al., 2011). Experiment 7 demonstrated that the Same-Test group suffered less from PI in the List 4 interim test than the Same-Math group, implying that interim testing can be positively used to prevent the accumulation of PI while studying additional material of the same type. For instance, after meeting a group of people, an individual might self-test herself on their names to alleviate PI before meeting the next group.

In conclusion, the forward testing effect is transferable even when material types and test formats are changed from block to block and transfers from low- to high-level learning. Prior interim tests induce greater test expectancy and motivate people to exert more effort toward encoding new information. Moreover, prior tests also induce people to exert more effort to retrieve the subsequently studied information. Instructors and learners are encouraged to administer interim tests during a study phase to facilitate subsequent learning of new information regardless of whether the material types and test formats are changed or not.

CHAPTER SIX: THE FORWARD BENEFITS OF INTERIM TESTING GENERALIZE TO OLDER ADULTS

Learning and memory deficits are common complaints among older adults. It is well-established that working memory (Whitebourne & Whitebourne, 2014), episodic memory (Old & Naveh-Benjamin, 2008), remote memory (Piolino, Desgranges, Benali, & Eustache, 2002), and prospective memory (Henry, MacLeod, Phillips, & Crawford, 2004) decline systematically as a function of age (for a review, see Lin & Fergus, 2008). How to mitigate older adults' learning and memory deficits is an important challenge. This chapter aims to explore whether interpolated testing can be employed as a remedial technique to facilitate older adults' learning and memory of new information.

As discussed in Chapter 1, different mechanisms may underlie the forward testing effects on single item learning and learning of complex materials (such as lecture videos and text passages). Szpunar et al. (2008) assumed that the forward testing effect on single item learning is largely driven by release from PI: Interim testing induces context changes, facilitates segregation of different learning events, and reduces the build-up of PI, producing better retrieval of new information. However, Wissman et al. (2011) noted that the release of PI mechanism contributes minimally to the forward testing effect on learning of complex materials because PI is less prevalent in complex materials. Instead, they proposed that the forward benefits of interim testing on learning of complex materials result from activation facilitation: Interim testing of studied information boosts retention of studied information (Roediger & Karpicke, 2006a), and better-remembered information (which is more retrievable and mentally activated) facilitates comprehension of new (related) information (Bransford & Johnson, 1972). Additionally, Szpunar et al. (2013) and Jing et al. (2016) found that interim tests while watching lecture videos substantially reduce participants' task-

irrelevant mind wandering and sustain learning engagement, implying that the learning engagement mechanism also underlies the forward testing effect on learning of complex materials.

In summary, different mechanisms may underlie the forward testing effects on learning of single items and of complex materials. This chapter explores whether the forward testing effects on such materials generalize to older adults. Specifically, Experiments 10 and 11 explored the forward testing effect on single item learning and Experiment 12 explored this effect on the learning of complex materials (lecture video).

Would we expect the forward testing effect to generalize to older adults?

It is reasonable to expect an affirmative answer to the question of whether the forward testing effect on learning of single items generalizes to older adults because this effect is substantial and robust. Furthermore, Pastötter et al. (2013) demonstrated that the effect occurs in patients with traumatic brain injury (TBI), who have deficits in remembering recent events. However, it is important to note that there are a few reasons to suspect that the effect might not generalize to older adults.

Aslan and Bäuml (2015) explored the forward testing effect on children's learning of single items. They instructed adults, older children (average age = 8.8 years), and younger children (average age = 6.7 years) to study four 6-word lists. Aslan and Bäuml obtained a forward testing effect among adults and older children, but this effect did not generalize to younger children. Aslan and Bäuml also observed that interim tests substantially prevented the build-up of PI for adults and older children, but not for younger children. Aslan and Bäuml explained that the absence of the forward testing effect in younger children's learning of single items might be caused by the facts that they have difficulty in inhibiting PI and that interim tests hence fail to attenuate the accumulation of PI for them. Similar to younger

children, older adults are more susceptible to PI than young adults (Ikier & Hasher, 2006; Ikier et al., 2008). Older adults commonly have difficulty dealing with interference, which is attributable to two psychological factors: encoding deficits (older adults are less able to prevent non-target information from entering memory) and retrieval deficits (they have difficulty in inhibiting non-target information while retrieving target information). Therefore, it is reasonable to assume that the forward testing effect in the learning of single items may not generalize to older adults because of their deficits in inhibiting PI.

As discussed above, the forward testing effect on learning of complex materials in younger adults might be driven by the facts that they frequently experience mind wandering across a study period and that interpolated tests reduce task-irrelevant mind wandering and enhance learning engagement. Recent research has demonstrated that older adults, by comparison with young adults, tend to experience less mind wandering while performing a range of tasks, such as reading text passages (Krawietz, Tamplin, & Radvansky, 2012), performing the Sustained Attention to Response Task (SART; Jackson & Balota, 2012), and executing the Operation Span Task (OSPAN; Jordano & Touron, 2017). Hence, it is reasonable to suspect that the forward testing effect on learning of complex materials may not generalize to older adults because they do not experience as much mind wandering as younger adults.

In summary, there are reasons to expect that interpolated testing can be employed as an efficient remedial technique to enhance older adults' learning and retrieval of new information. However, there are also grounds for believing that the forward testing effect will not generalize to older adults. Given that it is difficult to derive clear *a priori* predictions, this chapter is motivated to explore this important question.

Experiment 10

Method

Participants

Experiment 10 was an online experiment and participants were recruited from Prolific Academic (https://www.prolific.ac/). Participation requirements were restricted to: 1) First language = English; 2) Aged ≥ 60; 3) Nationality = UK; 4) Current resident = UK.

The sample size was determined according to the effect size (Cohen's d = 1.12) in Experiment 4. The required sample size to observe a significant (α = .05; power = .80) forward testing effect on learning of single items is 14 participants per group. To permit counterbalancing of the list sequence (see below), the final sample size was set to 15 participants in each group. Thirty older adults (average age = 68.43 years, *SD* = 3.97; 17 females) were recruited and randomly divided into two (Interim Test/Interim Math) groups. Ten participants' highest education level was graduate degree, 10 was undergraduate degree, eight was college/A-level, one was secondary school/GCSE, and one had no formal qualification.

Materials, design, and procedure

The principal stimuli were five 18-word lists, taken from Experiment 4. The list sequence was counterbalanced across participants using a Latin square design. Participants were informed that they would study five lists of words. Their task was to remember as many words as they could in preparation for a final cumulative test, in which they would be asked to recall as many words from all five lists as they could. They were also informed that, following studying each individual list and solving math problems for 1 min, the computer would decide at random whether to give them a short interim test. If it did, they needed to recall the words from the just-studied list. If not, they would continue solving math problems

for another 1 min. In fact, the interim test decisions were predetermined. The Interim Test group took an interim test after studying each list. By contrast, the Interim Math group continued solving math problems on Lists 1-4 and took an interim test on List 5.

In each individual list's study phase, 18 words were presented one-by-one in a random order, at the center of the screen, for 4 sec each, and each word was preceded by a fixation cross (presented for 500ms) as an interstimulus interval (ISI). Following studying each list, both groups solved math problems (e.g., *45 + 62 = ?*) for 1 min. In the interim tests, participants were instructed to recall as many words from the just-studied list as they could. Following the completion of List 5, both groups took a final cumulative test in which they were required to recall as many words from all five lists as they could. There was no time limit and no corrective feedback in any of the tests. At the beginning of the experiment, participants were warned not to take notes while learning. At the end of the experiment, they were asked whether they had made notes to help their memory and all reported they had not. The experiment lasted about 25 min.

Results

Figure 6.1A presents interim test recall for both groups. For the Interim Test group, a repeated measures analysis of variance (ANOVA) with List (1-5) as a within-subjects variable revealed that interim test recall linearly decreased across lists and the linear reduction trend was marginally significant, $F(1, 14) = 3.80$, $p = .07$, $\eta_p^2 = .21$. This result indicates that interpolated tests do not completely prevent the reduction in efficiency of older adults' learning and retrieval of new single items across lists.

Of critical interest is the difference in List 5 test recall between groups. The Interim Test group ($M = 5.87$, $SD = 1.21$) recalled about twice many List 5 words as the Interim Math group ($M = 3.20$, $SD = 2.01$), difference = 2.67 words, 95% CI = [1.02, 4.32], Cohen's d =

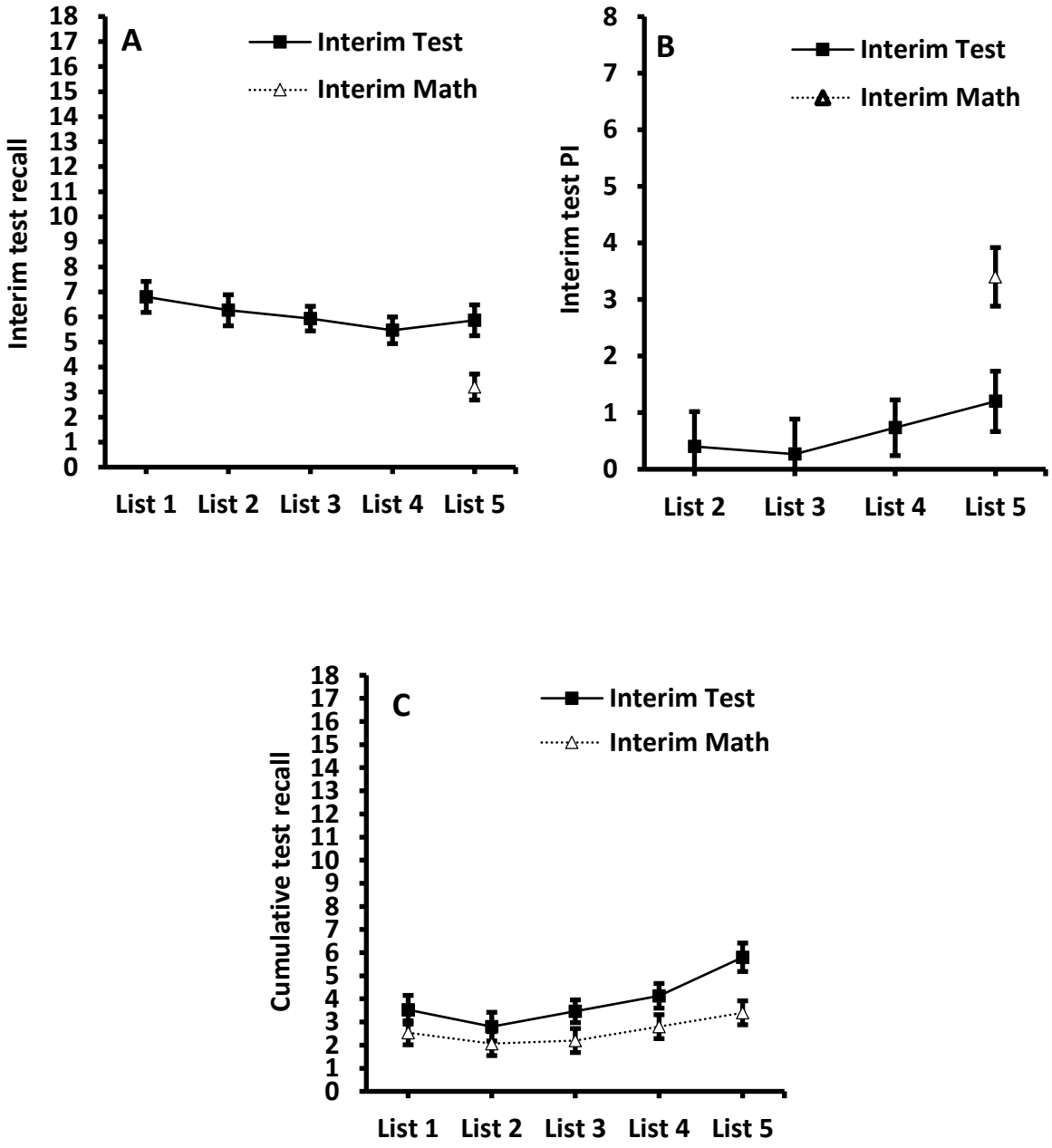

Figure 6.1: Experiment 10. Panel A: Interim test recall across five lists. Panel B: PI across the interim tests following Lists 2-5. Panel C: Cumulative test recall across five lists. Error bars represent ± 1 standard error.

1.21, indicating that interpolated tests facilitate older adults' learning and retrieval of new single items.

Figure 6.1B shows PI (i.e., incorrectly intruding words from prior lists while recalling words from the just-studied list) in the List 2-5 interim tests. For the Interim Test group, a repeated measures ANOVA, with List (2-5) as a within-subjects variable, showed that PI linearly increased across lists, $F(1, 14) = 8.15$, $p = .01$, $\eta_p^2 = .37$, indicating that interim tests do not completely prevent the build-up of PI across lists for older adults. Importantly, the Interim Math group ($M = 3.40$, $SD = 1.59$) experienced about three times as much PI as the Interim Test group ($M = 1.20$, $SD = 1.21$) in the List 5 test, difference = 2.20 words, 95% CI = [1.14, 3.20], Cohen's $d = 1.56$, indicating that interim tests (at least partially) prevent the build-up of PI for older adults.

Figure 6.1C depicts the cumulative test recall. The Interim Test group ($M = 19.73$, $SD = 7.24$) recalled more List 1-5 words than the Interim Math group ($M = 13.00$, $SD = 6.95$), difference = 6.73 words, 95% CI = [1.43, 12.04], Cohen's $d = 0.95$.

Summary

Experiment 10 revealed that, although interim tests do not completely prevent the build-up of PI, they substantially enhance learning and retrieval of new single items for older adults.

Experiment 11

Experiment 11 was designed to conceptually replicate Experiment 10's key findings, but with three modifications. First, Experiment 10 showed that interim testing, by comparison with solving math problems, enhances older adults' learning and retrieval of new single items. Experiment 11 compared interim testing with a restudying rather than a passive control group. Specifically, the Interim Math group was replaced by an Interim Restudy group, which

restudied Lists 1-4 and took an interim test on List 5. Second, participants' basic cognitive abilities were measured to ensure that the forward testing effect is produced by prior interim tests rather than by other uncontrolled factors. Third, Experiment 11 aimed to examine the forward testing effect in more naturalistic conditions. Specifically, older adults were instructed to remember common supermarket products.

Method

Participants

The sample size was determined according to Experiment 10's effect size (Cohen's *d* = 1.21). The required sample size to observe a significant (α = .05; power = .80) forward testing effect in the learning of single items is 12 participants in each group. In order to counterbalance list sequence, the sample size was set to 15 participants per group. Thirty-three older (aged over 60) adults were recruited from the local community of Wuhan University, China. One participant terminated the experiment prematurely, leaving a final sample of 32 participants, 24 females. They were randomly assigned to two groups, with 16 in the Interim Test group and 16 in the Interim Restudy group.[13] All participants' first language was Chinese. They were individually tested in a soundproofed room and received 50 RMB (about £6) as compensation. Wuhan University (WHU) Department of Psychology provided ethical approval for Experiments 11 and 12. [14]

Materials, design, and procedure

[13] Because of over-recruitment, the final data came from 32 participants, slightly larger than the pre-planned number (30). Excluding the final participant's data in each group did not materially change the results.

[14] Experiments 10-12 were conducted as an international collaboration project. Hence, Experiment 10 was conducted in the UK and Experiments 11 and 12 in China.

Fifty common supermarket item names (e.g., *apple, toothbrush, shampoo*) were taken from a Chinese shopping website (https://wenku.baidu.com/view/9ef45d97915f804d2b16c1f0.html). These items were randomly divided into five lists, with 10 items in each list. Each item name consisted of two or three Chinese characters. List sequence was counterbalanced across participants using a Latin square design.

At the beginning of the experiment, participants reported their age, gender, and education level. They were then asked to report their general mental and physical health status on a rating scale ranging from 1 (not well at all) to 5 (very well).

Next, participants were instructed to imagine that they were going to a supermarket and that the computer would show them five lists of items that they needed to buy. Their task was to remember as many items as they could. Other instructions and aspects of the experimental procedure were the same as in Experiment 10 with the following exceptions. The Interim Restudy group restudied Lists 1-4. Following studying each of Lists 1-4 and solving some math problems for 1 min, the items from the just-studied list were presented one-by-one in a new random order, for 4 sec each, for participants to restudy.

After finishing the final cumulative test, participants' vocabulary ability, working memory capacity (WMC), and processing speed were measured. The experiment lasted about 60 min.

Vocabulary task

The vocabulary task was taken from the Wechsler Adult Intelligence Scale (WAIS; Wechsler, 1955). Twenty Chinese words were selected from the Chinese version of the WAIS (Dai & Gong, 1987). The experimenter read the words one-by-one in a random order. Participants were required to identify the words in a table and explain their meaning. If a

given word was identified correctly, it received 1 point. If the word-meaning explanation was correct, that word received another point. Hence, the total score ranged from 0 to 40.

Digit Span task

WMC was measured through a digit span task, taken from the WAIS Scale (Wechsler, 1955), in which the experimenter read a sequence of digits and participants were required to repeat them either in the order in which they were read out or in the reverse order (for a detailed description of this task, see GrÉGoire & Van Der Linden, 1997).

Processing speed task

In the processing speed task, participants were asked to decide, as quickly and accurately as they could, whether two lines of digits were completely identical (Wang, Shen, Peng, Tang, & Zhang, 2005). In each trial, two lines of digits, with nine digits in each, were shown alongside each other on screen. There were 18 trials in total, with nine same trials and nine different trials. For the different trials, eight digits were identical and one was different. Participants were given two practice trials before the main task. Both response times (RTs) and identification accuracy were measured.

Results

There was no significant difference between the two groups' age, educational level, health rating, WMC, vocabulary ability, or processing speed (both response accuracy and speed; see details in Table 6.1).

Figure 6.2A depicts interim test recall. A repeated measures ANOVA, with List (1-5) as a within-subjects variable, showed that the Interim Test group's recall linearly decreased across lists, $F(1, 15) = 17.87$, $p = .001$, $\eta_p^2 = .56$, again indicating that interim tests do not completely prevent older adults' learning and retrieval of new single items from decreasing

Table 6.1: Demographic and basic cognitive ability results in Experiments 11 and 12

Groups and differences	Age (year)	Education (year)	Health (1-5)	Vocabulary (out of 40)	WMC	Processing speed accuracy (%)	Processing speed median RTs (sec)
Experiment 11							
Interim Test: *M* (*SD*)	66.47 (4.93)	9.93 (3.28)	2.73 (0.46)	36.87 (3.31)	5.73 (0.98)	92.96 (7.99)	6.54 (1.11)
Interim Restudy: *M* (*SD*)	65.63 (4.24)	9.25 (2.29)	2.75 (0.77)	35.94 (3.15)	5.40 (0.95)	89.58 (10.32)	6.29 (1.81)
Difference: *t* (*p*) values	0.51 (.61)	0.68 (.51)	-0.07 (.94)	0.80 (.43)	0.94 (.35)	1.02 (.32)	0.46 (.65)
Experiment 12							
Interim Test: *M* (*SD*)	67.16 (6.44)	10.92 (3.11)	3.08 (0.64)	37.08 (3.21)	5.84 (1.07)	96.22 (5.25)	6.37 (2.13)
Interim Restudy: *M* (*SD*)	66.96 (6.20)	11.04 (3.11)	3.21 (0.83)	36.21 (4.23)	5.73 (1.22)	96.30 (4.54)	5.73 (1.22)
Difference: *t* (*p*) values	0.11 (.91)	-0.14 (.89)	-0.61 (.55)	0.81 (.42)	0.34 (.74)	-0.05 (.96)	0.48 (.64)

Note: WMC = working memory capacity. Mean (*SD*) for demographic and basic cognitive ability results across experiments. The differences (*t* and *p* values) between groups are also reported.

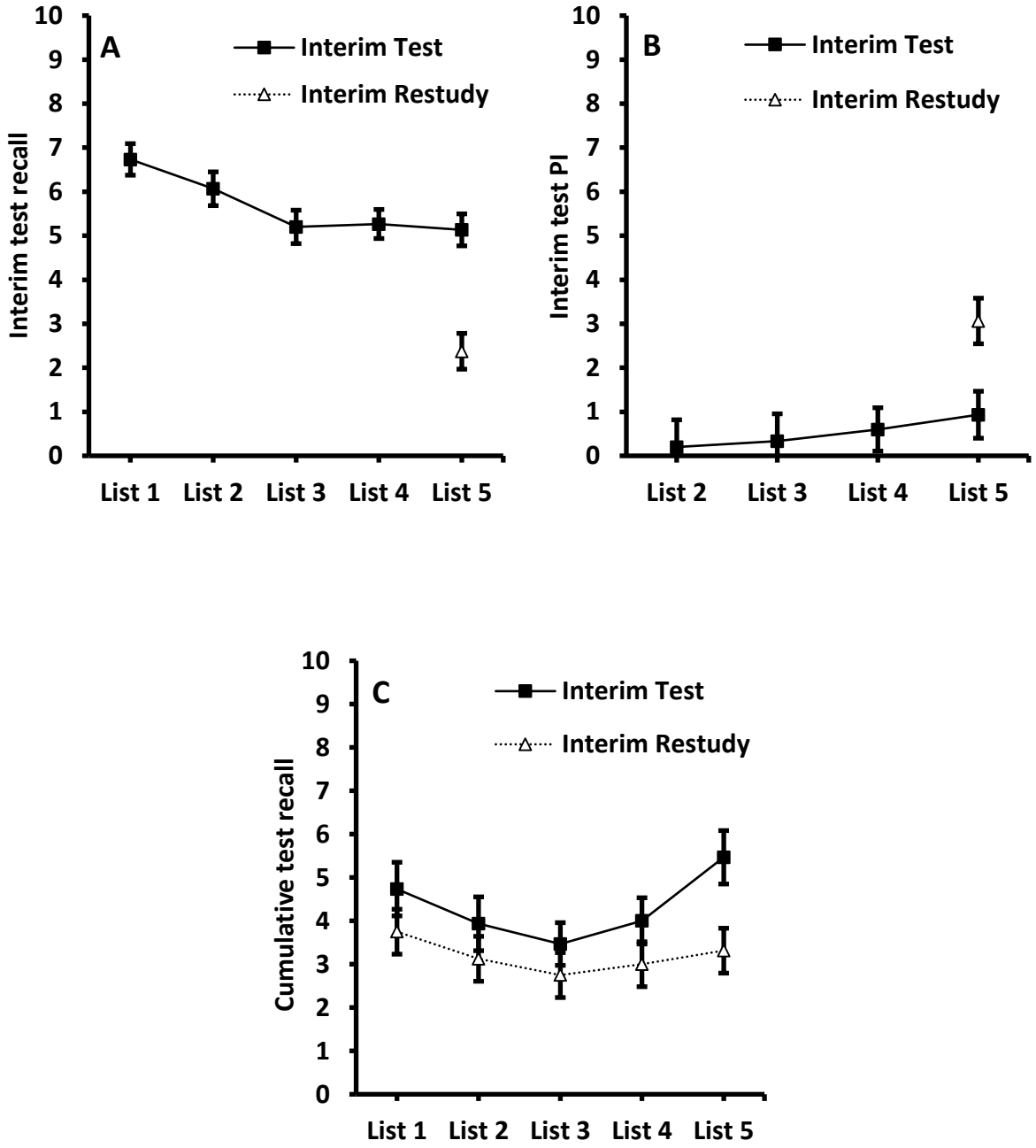

Figure 6.2: Experiment 11. Panel A: Interim test recall across five lists. Panel B: PI across List 2-5 interim tests. Panel C: Cumulative test recall across five lists. Error bars represent ± 1 standard error.

across successive lists. In the List 5 test, the Interim Test group (M = 5.13, SD = 1.41) recalled about twice as many List 5 words as the Interim Restudy group (M = 2.38, SD = 1.63), difference = 2.76 items, 95% CI = [1.64, 3.88], Cohen's d = 1.80.

Figure 6.2B depicts PI in the List 2-5 interim tests. A repeated measures ANOVA showed that, for the Interim Test group, PI linearly increased across the List 2-5 interim tests, $F(1, 14) = 20.70$, $p < .001$, $\eta_p^2 = .60$. In the List 5 test, the Interim Restudy group (M = 3.06, SD = 2.11) experienced about three times as much PI as the Interim Test group (M = 0.93, SD = 0.96), difference = 2.13 items, 95% CI = [0.91, 3.35], Cohen's d = 1.30, again indicating that interim tests substantially reduce the build-up of PI.

Figure 6.2C depicts the cumulative test recall. The Interim Test group (M = 21.80, SD = 4.86) recalled more List 1-5 words than the Interim Restudy group (M = 16.13, SD = 5.26), difference = 5.68 items, 95% CI = [1.95, 9.45], Cohen's d = 1.12.

Summary

Experiment 11 conceptually replicated Experiment 10's key findings, using more naturalistic stimuli relevant to a daily life situation, indicating that the forward benefits of interim testing on single item learning generalize to older adults.

Experiment 12

Experiments 10 and 11 demonstrated that interpolated testing enhances older adults' learning and retrieval of new single items. Experiment 12 tested whether the forward testing effect generalizes to older adults' learning of complex materials (lecture videos).

Method

Participants

Participants were recruited from a *Positive Psychology for Older Adults* course in a Chinese senior community college. The experiment was advertised at the end of a lecture and 52 older (aged over 60) adults volunteered to participate. Three of them were excluded because they did not return for testing of their basic cognitive abilities, leaving data from 49 participants (38 females). They were randomly divided into two groups, with 25 in the Interim Test group and 24 in the Interim Restudy group. According to the effect size from Szpunar et al.'s (2013) Experiment 2 (Cohen's d = 1.06), with 24 participants in each group the power to obtain a significant (α = .05) forward testing effect is 0.95. No participant had taken part in a previous experiment on the forward testing effect. Their first language was Chinese and all instructions and stimuli were in Chinese. They received 60 RMB (about £7) for compensation and were tested individually in a soundproofed room.

Materials, design, and procedure

The principal stimulus was a lecture video, concerning *Efficient Communication between Doctors and Patients*, taken from the NetEase Open Course website (available at http://open.163.com/movie/2015/2/S/8/MAH8PH40M_MAKPPCAS8.html). The lecture was split into four segments, each lasting about 5 min. A short summary, consisting of 15 short sentences, was created for each segment. In each sentence, there was a critical word which was probed in the tests.

At the beginning of the experiment, participants reported their demographic information. They were then instructed to study a four-segment lecture video. They were informed that they would take a final cumulative test on all four segments following the completion of Segment 4, and the computer would randomly decide whether or not to give them a short test after studying each segment and solving some math problems for 30 sec. If it did, they would take a test on the just-studied segment. If it did not, they would restudy that

segment. In fact, the Interim Test group took an interim test following each segment whereas the Interim Restudy group restudied Segments 1-3 and only took an interim test on Segment 4.

In the interim tests, a short summary of the just studied segment was presented on screen. It consisted of 15 sentences, with a blank in each sentence (e.g., *Doctors think that they should first ____ themselves in medical disputes*). Participants were required to type their answers into the corresponding blanks. They had unlimited time to answer each question and were allowed to leave a space empty if they did know the correct answer. Following each interim test, participants received corrective feedback. The summary was shown with the correct answer underlined and in red (e.g., *Doctors think that they should first protect themselves in medical disputes*) and participants had 120 sec to study the feedback. By contrast, the Interim Restudy group restudied Segments 1-3. Specifically, the short summary was presented for 300 sec for them to restudy. The 15 critical targets were underlined and in red, informing the Interim Restudy group which words would be tested at the cumulative test.

Following the completion of Segment 4, both groups took a cumulative test comprising 20 fill-in-the-blank questions, with five questions chosen from each of Segments 1-4. No corrective feedback was provided. Only 20 questions were included instead of all 60 because some participants in a pilot study reported fatigue at the length of the test. About 24 h later, they returned to complete the cognitive ability tests, which were identical to those administered in Experiment 11. The experiment lasted about 70 min in total.

Results

Demographic information and basic cognitive ability results are reported in Table 6.1. There was no significant difference between groups in age, education, health rating, vocabulary ability, WM, or processing speed (both accuracy and median RTs).

Two research assistants, who were naïve to the experimental hypothesis, scored recall performance. One assistant scored all participants' interim and cumulative test recall. Another only scored Segment 4 test responses. 95.9% of their scores (i.e., scores of the Segment 4 test) were identical, and the discrepant scores were resolved by discussion.

In the Segment 1-3 interim tests, the Interim Test group's recall was 9.00 (*SD* = 2.42) items, 9.08 (*SD* = 3.09), and 7.76 (*SD* = 2.52), respectively out of 15. Of critical interest is the difference in the Segment 4 test recall between groups (see Figure 6.3A). The Interim Test group (*M* = 6.32; *SD* = 3.11) recalled about twice as many items correctly as the Interim Restudy group (*M* = 3.50; *SD* = 2.17), difference = 2.82 items, 95% CI = [1.28, 4.36], Cohen's *d* = 1.05, revealing that interpolated testing enhances older adults' learning of new complex materials. As shown in Figure 6.3B, the Interim Test group (*M* = 15.84; *SD* = 2.29) also outperformed the Interim Restudy group (*M* = 12.21; *SD* = 3.87) in the cumulative test, difference = 3.63 items, 95% CI = [1.81, 5.45], Cohen's *d* = 1.14.

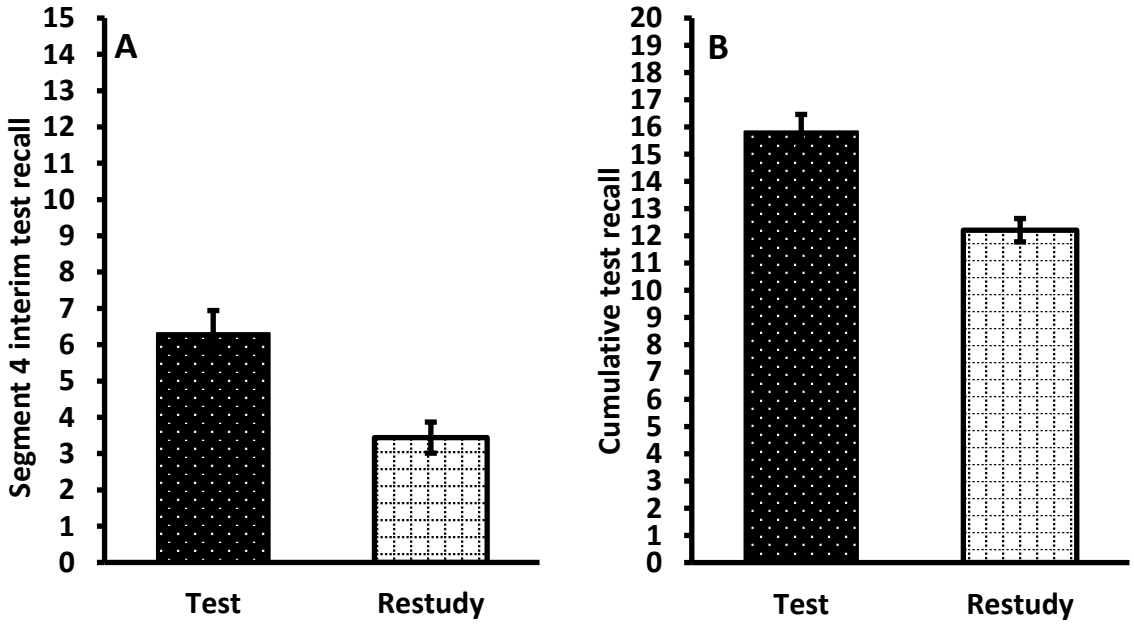

Figure 6.3: Experiment 12. Panel A: Segment 4 interim test recall. Panel B: Cumulative test recall. Error bars represent ± 1 standard error.

Summary

Experiment 12 demonstrated that the forward testing effect in older adults extends to the learning of complex materials.

Discussion

This chapter explored whether interim testing aids older adults to learn and retrieve new information. Experiments 10 and 11 investigated the forward testing effect on older adults' learning of single items and Experiment 12 explored the effect on learning of complex materials. Experiment 10, employing five lists of words, showed that interim testing significantly reduces the build-up of PI and enhances learning and recall of new single items. Experiment 11, employing five lists of common supermarket items, conceptually replicated Experiment 10's key findings. Both experiments found that participants' recall linearly decreased and PI linearly increased across lists, providing additional evidence supporting the release from PI theory to account for the forward testing effect on single item learning.

In summary, this chapter shows that interim testing enhances older adult's learning and retrieval of new single items, although it did not completely prevent the decrease of recall and the build-up of PI across lists. In addition, interim tests aid older adults' learning of lecture videos. Clearly, the present results do not tell us whether the effect is as large amongst older as amongst younger adults. Future research addressing this question would be valuable.

CHAPTER SEVEN: GENERAL DISCUSSION

This thesis has explored several important aspects of the forward testing effect: self-regulated study time allocation (Experiments 1 and 2); metamemory monitoring (Experiments 3, 4, and 6); inductive learning (Experiments 5 and 6); transfer (Experiments 7-9); and older adults' learning of single items and complex materials (Experiments 10-12).

This chapter aims to summarize the main empirical findings of Experiments 1-12, offer an overview of the current state of knowledge more generally, and discuss the practical implications for optimizing learning and teaching in educational settings. The possible negative effects of interim testing on learning of new information and how to mitigate such effects are also discussed. Finally, this chapter provides some suggestions for future research to further investigate aspects of this important effect that are currently poorly understood.

Figure 7.1 summarizes the characteristics [participants, stimuli, learning procedures, and effect types (the classic forward testing effect and the transfer effect)] of Experiments 1-12 and the effect sizes of the forward testing effects observed in these experiments. A random-effects meta-analysis, which extracted the data from 653 participants in the 12 experiments, found that the effect size (Cohen's *d*) was 0.95, 95% CI = [0.77, 1.12], indicating the robustness of the forward testing effect. The heterogeneity of the effect sizes was non-significant, $Q(12) = 12.36$, $p = .42$, precluding any analysis of moderator variables. Below this chapter will separately discuss the forward testing effects for different types of learning and illustrate its educational implications.

Single item learning

Experiments 4, 10, and 11 consistently observed that interim tests enhance learning and retrieval of new single items, a finding repeatedly demonstrated by previous studies using

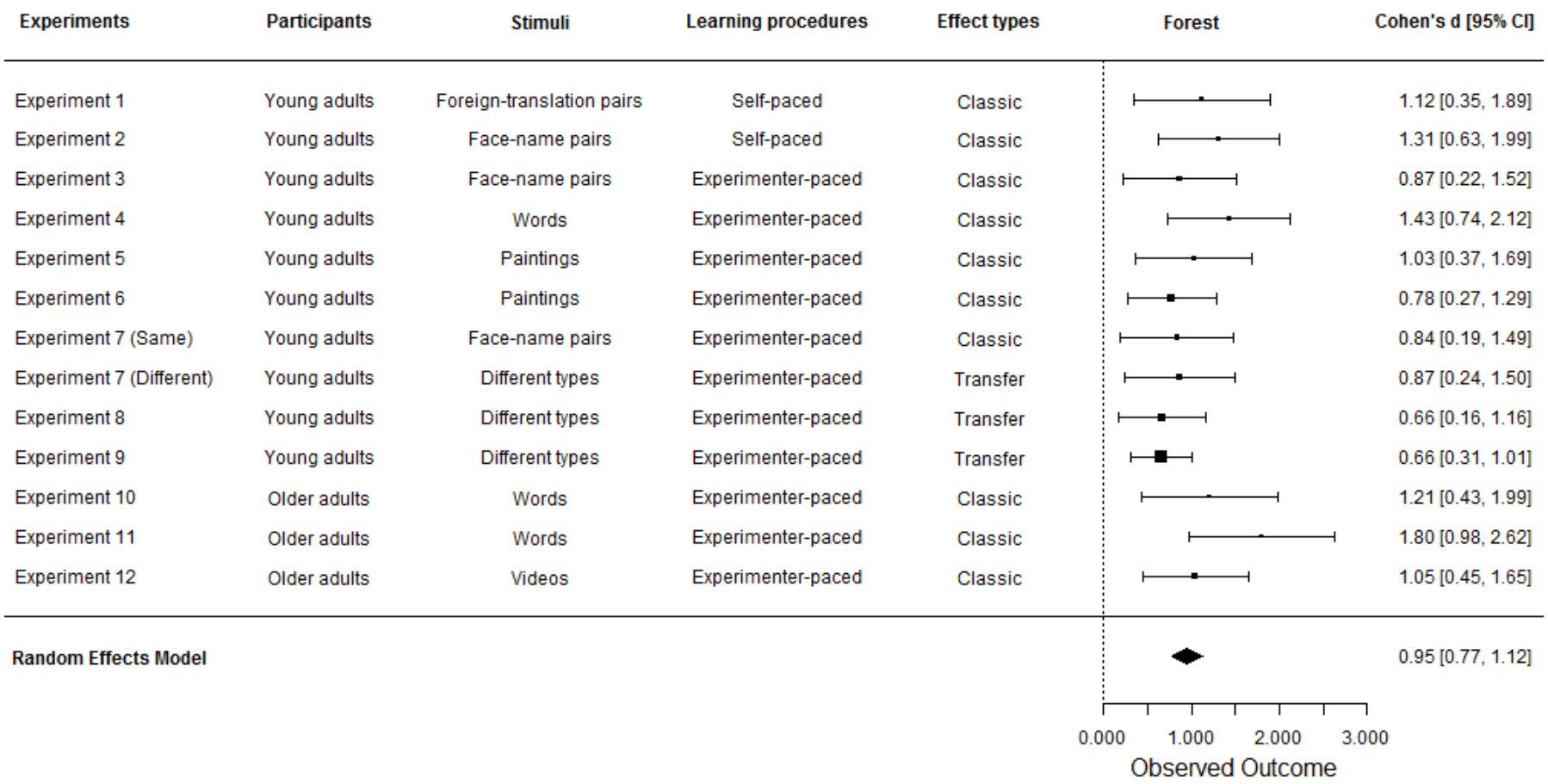

Figure 7.1: Forest plot summarizing the effect sizes and characteristics (participants, stimuli, learning procedures, and effect types) of Experiments 1-12. Error bars represent 95% CI.

word (Aslan & Bäuml, 2015; Bäuml & Kliegl, 2013; Nunes & Weinstein, 2012; Pastötter, Schicker, Niedernhuber, & Bäuml, 2011; Pierce, Gallo, & McCain, in press; Weinstein, Gilmore, Szpunar, & McDermott, 2014) and picture (Pastötter, Weber, & Bäuml, 2013) lists. At the same time, all three experiments observed that interim testing substantially reduces the build-up of PI across lists/blocks. These results clearly reveal that interim testing of studied single items, compared with restudying or no interim testing (distractor task), enhances learning and retrieval of new single items and prevents the accumulation of PI.

It is however unclear whether the effect of interim testing on release from PI will endure over the long term. To our knowledge, in all previous studies, the effect of interim testing on release from PI was explored with a short retention interval (i.e., the interval between studying the final list and taking the interim test ranged from 0-25 min). It has not been explored whether the release from PI, induced by interim testing, is long lasting. This question awaits exploration in future research.

Paired-associate learning

Experiments 1, 2, 3, 7, and 8 showed that administering interim tests facilitates learning and retrieval of Euskara-English word pairs, face-name pairs, and face-profession pairs. These results, consistent with the findings from other studies (Cho, Neely, Crocco, & Vitrano, 2016; Weinstein, McDermott, & Szpunar, 2011), clearly reveal that interim testing of studied paired-associates, compared to no interim testing or restudying, enhances learning and retrieval of new paired-associates. In addition, Experiments 1, 2, 3, and 7 again demonstrated that interim testing reduces the accumulation of PI.

Learning of complex materials

The forward testing effect on learning of complex materials was observed in Experiment 12, in which participants studied a four-segment lecture video. The forward

testing effect on the learning of lecture video content has also been reported by Szpunar, Jing, and Schacter (2014) and Yue, Soderstrom, and Bjork (2015), and the effect on learning of text passages was reported by Healy, Jones, Lalchandani, and Tack (in press) and Zhou, Yang, Cheng, Ma, and Zhao (2015). Going beyond previous studies, which explored this effect on young adults' learning of complex materials, Experiment 12 is the first observing this effect in older adults. Interim testing not only enhances memorization of specific content but also boosts information integration (Jing, Szpunar, & Schacter, 2016) and comprehension of complex materials (Zhou et al., 2015).

Inductive learning

Previous studies explored the forward testing effect on the learning of specific items (e.g., words, pictures, paired-associates, lecture videos, and passages). Experiments 7 and 8 documented the effect on inductive learning, and a similar effect was also recently reported by Lee and Ahn (in press). Lee and Ahn instructed three (Interim Test/Interim Restudy/Interim Distractor) groups of participants to study 36 paintings, comprising six from each of six artists, in Block 1. Then the Interim Test group took a cued recall test on Block 1 paintings, in which the same 36 paintings were presented one by one in a random order and participants were asked to recall the corresponding artists' names. The Interim Restudy group restudied the 36 paintings and the Interim Distractor group solved some math problems. In Block 2, all three groups studied 36 new paintings from another six artists. Next, they took a classification test, in which 48 completely new paintings, comprising four from each of the 12 studied artists, were presented one by one in a random order, with the 12 artists' names presented simultaneously with each painting. Participants were instructed to select which was the correct artist for a given painting. The results showed that the Interim Test group substantially outperformed the Interim Restudy and Interim Distractor groups, revealing a forward testing effect on inductive learning.

Inductive learning is a key element of how humans learn and understand the world and a key component of formal education. For example, fine art/history of art students are required to learn the painting styles of different artists; medical students are required to learn how to diagnose different diseases; students of linguistics have to learn the rules of a language; airport security screeners are required to learn how to detect threatening items by examining x-ray images. As illustrated by Experiments 5 and 6 and Lee and Ahn (in press), interim testing during studying is an effective strategy for improving inductive learning.

Self-regulated learning

Previous studies explored the forward testing effect in experimenter- or instructor-paced situations. But of course the pace of studying is often self-determined. Experiments 1 and 2 investigated the forward testing effect in a self-paced situation and showed that self-determined study time in the Interim Math group systematically decreased across lists whereas the decrease of study time was prevented by interim tests in the Interim Test group. When participants could choose how much time to devote to learning, their study time gradually dropped across successive lists in the absence of interim tests. However, when each list was followed by a test, participants maintained their list-by-list study time. These results, therefore, reveal that interim testing motivates people to commit more effort (study time) toward encoding new information in self-regulated learning, and shows that a simple intervention can motivate learners to devote more time to studying.

With the development of online courses and learning aids, self-regulated learning is becoming more and more common outside of the formal classroom (Bjork, Dunlosky, & Kornell, 2013). Yet we are far from being sophisticated learners (Bjork et al., 2013) and our self-regulated learning is often inadequate to achieve full mastery of the material we are studying (Kornell & Bjork, 2008). Administering interim tests during studying is a potent

strategy to promote and sustain the effectiveness of self-regulated learning across a learning phase.

Transferability of the forward testing effect

Students typically study different types of information from class to class. A biology class may be followed by a history class, for instance. It is therefore important to explore the transferability of the forward testing effect – whether interim testing of studied information in one domain can enhance learning and retention of new information in a different domain. Experiments 7-9 explored this important issue. Experiment 7 demonstrated the transferability of the effect; Experiment 8 showed that it transfers even when material types and test formats are switched from block to block; Experiment 9 observed transfer from low- (verbal facts) to high-level (visual concepts) learning. These results consistently reveal the robust transferability of this effect.

Because Experiments 7-9 did not include groups which studied materials from the same domain in each block, it is unknown whether the forward testing effect is attenuated by a change in domain compared to no change in domain. Experiment 7 included two groups which studied material from the same domain in each block. However, the Swahili-English word pairs were easier to remember than the face-name pairs, so the experiment does not yield a clear conclusion. Hence, to what extent the standard forward testing effect is attenuated by a change of material or by a change of test format is currently unknown and must await future research.

Metacognitive awareness of the forward benefits of interim testing

Experiments 3 and 4 explored the forward testing effect on metacognitive monitoring. Both experiments observed that, although participants' JOLs decreased across lists in both groups, JOLs in the Interim Test group decreased much less than those in the No Interim Test

group. More importantly, the alignment between JOLs and recall performance in the final list interim test (i.e., the Interim Test groups offered higher JOLs on the final list and correspondingly performed better in the final list interim test) reveals people's metacognitive insight into the forward benefits of interim testing.

However, it is unknown whether this metacognitive insight into the forward testing effect is explicit (learners might appreciate that their learning and recall is enhanced *because* they took an earlier test) or implicit (it is possible that prior interim tests maintained the Interim Test group's effort in encoding subsequent new information, but they did not know explicitly that the reason they allocated more effort was because of the prior tests). 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. Future research is needed to explore whether this awareness is explicit or implicit.

Individual differences

Two previous studies have explored individual differences in the forward testing effect. Pastötter et al. (2013) found that the forward testing effect on single item learning generalizes to TBI individuals. Aslan and Bäuml (2015) found that the effect on single item learning generalizes to older children but not to younger children.

Going beyond previous studies, Experiments 10-12 explored the forward testing effects on older adults' single item learning and learning of complex materials. Experiments 10 and 11 showed that, although interim tests did not completely prevent the decrease of learning and retrieval of single items and could not completely eliminate the build-up of PI across lists, they did substantially facilitate learning and reduced the accumulation of PI across lists for older adults. Experiment 12 showed that interim tests aid older adults to learn

the content of lecture videos. These findings suggest that interim tests are effective techniques to mitigate older adults' learning and memory deficits.

As yet, it is still unknown whether interim testing can be used to mitigate memory deficits caused by conditions such as Alzheimer's disease, ADHD, and multiple sclerosis. Future research is encouraged to explore this. Aslan and Bäuml speculated that the absence of the forward testing effect in younger children's single item learning may result from their deficits in inhibition of PI. For younger children, the absence of the forward testing effect on single item learning does not presuppose the absence of this effect on learning of complex materials, as the activation facilitation and enhanced encoding effort mechanisms may play important roles for learning of complex materials whereas the release from PI mechanism is likely to play little role (Wissman, Rawson, & Pyc, 2011). Future research could profitably explore this issue.

Theoretical implications

Chapter 1 summarises eight theories proposed to account for the forward testing effect. These accounts are not mutually exclusive, most are at a preliminary stage of development, and few have been subjected to direct testing of their key predictions. The findings derived from the thesis bear some important theoretical implications.

As summarised in Table 1.1, some theories (encoding reset; activation facilitation; encoding strategy; test expectancy; failure-encoding-effort) assume that the main locus of the underlying mechanism(s) of the forward testing effect is at the encoding phase whereas others (release from PI; retrieval strategy; retrieval effort) assume the locus is at the retrieval phase. In Experiments 1 and 2, the forward testing effect was replicated when the encoding procedure was self-paced. More importantly, both experiments showed a decreasing slope of encoding time across lists in the absence of interim tests, 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. Both experiments showed evidence that this forward benefit of interim tests was associated with a reduction in the amount of PI experienced in the final list interim test. Collectively, these results suggest that the forward benefits of interim testing are attributable to both encoding (e.g., greater effort and deeper encoding) and retrieval (e.g., greater list discrimination and release from PI) processes.

Experiments 7-9 demonstrated the transfer of the forward testing effect. The release-from-PI, encoding-reset, and activation-facilitation theories have difficulty explaining this transfer. To illustrate, consider Experiment 9. First, a switch of material types led to no PI in the Block 4 interim test, and materials in Blocks 1-3 (text statements) and Block 4 (paintings) were from different domains and completely unrelated. Hence, the transfer could be accounted for by neither release from PI nor activation facilitation. Second, a switch of material types also induces substantial context changes between blocks, which should "reset" subsequent encoding (Ellis & Montague, 1973; Emery, Hale, & Myerson, 2008; Nunes & Weinstein, 2012). Therefore, the encoding reset theory is also unlikely to account for successful transfer. Third, given that materials in different blocks were from different domains and completely unrelated, the activation-facilitation mechanism should have contributed little to transfer.

The strategy-change mechanisms proposed by the encoding-strategy and retrieval-strategy theories might have contributed to the transfer findings in Experiment 7, because Swahili-English word pairs and face-name pairs were both paired-associates and the test formats were always cued-recall in all interim tests. However, these theories have difficulty explaining the transfer findings in Experiments 8 and 9, in which material types and test formats were both switched, because there is little reason to expect that participants would

have developed and adopted more effective encoding/retrieval strategies across prior blocks that would be applicable in the final block.

In contrast, the test expectancy theory readily explains the transfer findings in Experiments 8 and 9. Indeed, Experiments 8 and 9 consistently observed that prior interim tests induced participants to expect an interim test on the next list/block, and there were positive (albeit modest) correlations between test expectancy and test performance, supplying evidence supporting the test-expectancy theory. By measuring participants' RTs (an index of retrieval effort), Experiment 9 showed that interim tests motivate people to exert more effort toward retrieving the target information, supplying new evidence supporting the retrieval effort theory.

In summary, eight theories have been proposed to account for the forward testing effect. They need not be mutually exclusive, and some of them may operate in parallel in some situations and produce overlapping forward testing effects. Future studies are needed to further explore these possible mechanisms and investigate in which situations and for which materials the different mechanism(s) contribute to the forward testing effect.

Possible negative effects of interim testing

The thesis largely focuses on the facilitatory effects of interim testing on learning and retrieval of new information. However, recent research has shown that in some situations interim testing can lead to negative effects. Interim testing motivates people to apply more effort to the encoding of new materials. It has been suggested that, when the tested and new materials are presented together, the tested materials may "forcibly occupy" the encoding time and borrow the limited time available for studying new materials – the *borrowed time effect* (Davis & Chan, 2015).

As an illustration of this, Finn and Roediger (2013) asked two (Interim Test/Interim Restudy) groups of participants to study some face-name-profession associations. In the first encoding phase, both groups studied face-name pairs one-by-one, for 5 sec each. Following a short distractor task, the Interim Restudy group restudied all face-name pairs one-by-one, for 5 sec each, and immediately following the presentation of each face-name pair, the same face with its name and profession were presented simultaneously for 5 sec for participants to study. In contrast, following the distractor task, the Interim Test group took an interim test in which they were asked to recall each face's corresponding name, and immediately following this recall, the same face was shown with its corresponding name as corrective feedback for 2 sec. After that, the face with its name and profession were simultaneously presented on screen for 5 sec. Twenty-four hours later, both groups took a final test, in which participants were shown the faces one at a time and asked to recall each face's corresponding name and profession. The results showed that the Interim Test group correctly recalled more names than the Interim Restudy group, whereas the recall of professions showed the reverse pattern: The Interim Restudy group recalled more professions than the Interim Test group. In recent follow-up research, Davis and Chan (2015) proposed that the interim test on face-name pairs led the Interim Test group to continue to focus on learning the names when they were shown the face-name-profession pairs, as the prior interim test made them aware of the difficulty of remembering the face-name pairs. This focus on the learning of names "borrowed" encoding time and resources which were supposed to be spent on learning professions.

Davis and Chan (2015) showed that this negative effect can be completely reversed. In their Experiment 4, they separated the interim test on face-name pairs and the encoding of face-profession pairs. Following initial studying of the face-name pairs, an Interim Restudy group restudied all face-name pairs one-by-one, whereas an Interim Test group was shown the faces one-by-one and was asked to recall the names, and corrective feedback was given in

the interim test. Then both groups were asked to study the face-profession pairs, and in this phase, no names were shown alongside. At the end of this study phase, participants took a final test in which they were asked to recall the names and professions in response to the faces. The results showed that, in the final test, the Interim Test group recalled more professions than the Interim Restudy group. Hence the Davis and Chan (2015) study showed that separate presentation of tested and new information can not only eliminate the borrowed time effect (the finding that interim testing on face-name pairs impairs the learning of face-profession pairs when face-name-profession information was simultaneous) but can also induce a positive forward testing effect (interim testing on face-name pairs enhances the learning of face-profession pairs when the associations are separated). However, the positive backward testing effect for the face-name pairs observed by Finn and Roediger (2013) was also reversed: now testing impaired memory for the names relative to restudy. In a more recent study, Davis and colleagues (Davis, Chan, & Wilford, 2017) provided evidence suggesting that the impairment to new learning could be due to a task switching cost rather than to time borrowing, though the two accounts are not mutually exclusive.

Finn (2017) reported finding that, contrary to the predictions of the borrowed time hypothesis, giving participants unlimited time for test and review of feedback, thereby minimizing the need to borrow time, failed to eliminate retrieval-impaired learning of new information. While time borrowing may account for some of the data, it is therefore unlikely to be a complete explanation for retrieval-impaired learning of complementary associations (for other possible explanations about why interim testing may impair learning of new information see Davis et al., 2017; Finn, 2017; Finn & Roediger, 2013). Finn and colleagues' finding (that is, interim testing impairs the learning of new complementary information when tested and new complementary information is simultaneously presented) is intriguing and research into this phenomenon is at an early stage. Future research is encouraged to further

explore the putative mechanisms underlying the negative effects of interim testing and develop practical interventions to eliminate (or at least minimize) these negative effects.

Clearly, there will be many occasions in the classroom when an instructor asks students to recall a piece of studied information (e.g., A-B associates) as a prelude to introducing new, complementary information (e.g., A-C associates). Finn and colleagues' studies suggest that simultaneous presentation of tested and new information (e.g., A-B-C associations) following retrieval of studied information (e.g., A-B associates) can impair the learning of new information (e.g., A-C associates) but the mechanisms underlying this effect are not yet clearly delineated. The finding that impairment of new learning is greatest in situations characterized by frequent task switching suggests that it may not pose a serious problem in classroom situations, where such frequent switching is unlikely, but not enough is yet known about the boundary conditions of the effect to make firm recommendations.

Future research directions

Table 7.1 lists some directions for future research to explore. As discussed above, more work is needed to explore the mechanisms underlying the forward testing effect. The participants in previous research were largely restricted to college students, and future research is encouraged to explore the generalizability of the effect to other groups of individuals. For instance, individuals with Alzheimer's disease suffer from significant memory deficits (Hodges, Salmon, & Butters, 1992), and more work is needed to explore whether interim testing can be employed as a remedial technique to mitigate their memory deficits. Individuals with attention deficit hyperactivity disorder (ADHD) have difficulty in sustaining their attention on the learning task (DuPaul & Volpe, 2009) and they are more susceptible to mind wandering (Seli, Smallwood, Cheyne, & Smilek, 2015). Therefore, it is important to test the utility of interim testing in enhancing their learning engagement and learning outcomes.

Table 7.1: Future research directions for investigating the forward testing effect.

Suggested future research directions
1. Further investigation on the possible mechanisms underlying the forward testing effect.
2. Can the forward testing effect generalize to individuals with Alzheimer's disease or ADHD?
3. Can young children's learning of complex materials benefit from interim testing?
4. What brain networks are involved in the forward testing effect?
5. Long-term outcomes of the forward testing effect.
6. The appropriate level of test difficulty for sustaining learning efficiency.
7. How many interim tests be administered, and how frequently?
8. Testing the forward testing effect in the classroom.
9. Exploring the social issues associated with testing.

For younger children, the absence of the forward testing effect in single item learning might result from deficits in inhibiting PI (Aslan & Bäuml, 2015). However, the absence of this effect in single item learning does not mean that the effect will also be absent in the learning of complex materials (e.g., texts), as the activation facilitation and enhanced encoding effort mechanisms may play important roles for complex materials whereas the release from PI mechanism is likely to play little role (Wissman, et al., 2011). Future research could profitably explore whether younger children's learning of text passages can benefit from interim testing.

To date, little neuroscientific research has been conducted to explore the human brain networks involved in the forward testing effect (Schacter & Szpunar, 2015). Exploring its

neural underpinnings will enrich our knowledge of the underlying mechanisms. For instance, many previous event-related functional magnetic resonance imaging (fMRI) studies found that information integration during encoding (i.e., binding disparate information into a coherent representation) is associated with greater activation levels in anterior hippocampal regions (e.g., Jackson & Schacter, 2004; Wing, Marsh, & Cabeza, 2013). If interim testing activates these regions more extensively than restudying or doing nothing, such evidence will support the activation-facilitation theory, which hypothesizes that interim testing facilitates knowledge-integration (Jing, et al., 2016) and improves comprehension of new information (Wissman, et al., 2011). Future neuroscience research is encouraged to begin to explore the neural underpinnings of the forward testing effect.

To date, forward testing effect studies have focused on the short-term outcomes of the effect: the target (final) block's interim test was administered immediately following its learning phase. Only one recent study documented that the effect on single item learning lasts at least 25 min (Chan, Manley, Davis, & Szpunar, 2018). Future research could profitably explore the long-term (e.g., 24 hrs, one-week, etc.) outcomes of the effect. Exploring the effect over longer intervals is important because, from an educational perspective, the case for administering interim tests in the classroom and elsewhere depends on their benefits being long-lasting.

Another question concerning the forward testing effect, which has not been explored yet, is whether test difficulty moderates the magnitude of the effect. Based on the failure-encoding-effort and retrieval-effort theories, difficult tests, by comparison with easy tests, will induce more retrieval failures, which in turn will induce greater encoding and retrieval effort and produce large forward benefits of testing. However, more retrieval failures may induce higher test anxiety, which in turn lead to difficulty in concentrating (Tse & Pu, 2012)

and hence may yield smaller or even negative forward effects of testing. Therefore, it is important to explore the appropriate level of test difficulty for sustaining learning efficiency.

The backward testing effect has been repeatedly demonstrated in real classrooms but the forward testing effect has not been yet. Almost all of the previous forward testing effect studies were conducted in laboratory settings. The robust effect in the laboratory does not guarantee its generalization to the classroom because of the intrinsic differences between these settings. For instance, in all previous laboratory research (including the current thesis), participants were tested on all items contained in the just-studied block, but testing all information presented in a class is clearly impossible in the real classroom. Therefore, it is of critical importance to explore whether partial testing (i.e., testing on a subset of items) can induce a forward testing benefit. The successful transfer of the effect observed in Experiments 7-9 does not guarantee successful transfer in the classroom, wherein no experimental context exists to link different classes. It is important for future research to explore the transfer effect in the classroom.

Although testing has been identified as one of the most efficient techniques to enhance learning (Dunlosky et al., 2013), educators often prefer to minimize the use of tests in the classroom because administering and scoring them is time-consuming and demanding. Hence, this raises an important question: How frequently should we administer interim tests to sustain students' engagement and enhance their learning efficiency? In summary, many important aspects of the forward testing effect need to be explored in the classroom. Without a much deeper exploration of the effect, educational translation and exploitation will be hindered.

This thesis focuses on the beneficial effects of interim testing on learning and retrieval. Future research should move to explore the effects of testing on social issues, such

as cheating on tests, the effect of testing on learning motivation. Testing is regarded as a tool for assessing students' learning outcomes, and test performance is sometimes regarded as an indicator of teachers' teaching effectiveness (Stronge, Ward, & Grant, 2011). To raise test scores, some students, teachers, and administrators may be stimulated to cheat on tests. A famous example is the Atlanta school cheating scandal (https://en.wikipedia.org/wiki/Atlanta_Public_Schools_cheating_scandal), wherein, to avoid negative evaluations, 178 teachers and principals in the Atlanta Public Schools district cheated on the standardized tests administered by the state. It is likely however that forward testing effects occur even with low-stakes, anonymous quizzes in which the incentive to cheat should be low.

From a positive perspective, testing may induce competition amongst students and teachers, and stimulate them to study and learn more effectively (for a review of the effect of testing on learning motivation, see Harlen & Deakin Crick, 2003). However, from a negative perspective, poor test performance may induce frustration and feelings of inadequacy, which might in turn impair learning and teaching outcomes. In summary, there are many social issues associated with testing, and more attention should be paid to exploring them in future research.

Summary

Learners and educators sometimes simply regard testing as a tool for measuring one's learning status. Some educators even propose minimizing testing in the classroom as they think it is time-consuming (Roediger & Karpicke, 2006a) and scoring of tests is demanding. However, numerous previous studies have confirmed the reliability of the backward testing effect in laboratories and real classrooms, even with low-stakes quizzes (Roediger & Karpicke, 2006a). The experiments contained in this thesis have consistently demonstrated

the reliability of the forward testing effect across a variety of educational materials. Therefore, the forward and backward testing effects, jointly, make a strong case for learners and instructors to administer interim tests or quizzes during learning.

Interim testing can not only enhance learning and retention of new information but also prevent the build-up of PI (Szpunar, McDermott, & Roediger, 2008; Weinstein et al., 2011; Experiments 1-4). In real-world learning settings, students frequently suffer from PI. For example, in a geography class, students may need to master basic information (e.g., geography, culture, economy, demographics) about some European countries (e.g., Norway, Denmark, and Spain). Students may confuse information relating to different countries. Therefore, understanding how to prevent the build-up of PI is critical for instructors and learners in such situations. As Experiments 1-4 showed, interim testing significantly decreases the build-up of PI regardless of whether the learning is self- or instructor-paced. However, it is important to be cautious because it is unknown whether release from PI, induced by interim testing, is long lasting.

To summarize, interim testing is a powerful technique for optimizing the learning of new information. Studies using a variety of educational materials have shown that the forward testing effect is a robust phenomenon. Interim testing can be used to enhance the learning and retrieval of new single items, paired-associates, complex materials, and concepts (categories). It not only benefits memorization of specific content but also boosts information integration, producing superior knowledge organization. The forward testing effect is not limited to instructor-paced situations but also generalizes to self-paced ones; it is not limited to healthy individuals but also generalizes to individuals with brain injury, older children, and older adults; it is not limited to the same type of material but is also transferable to different types of material (and different test formats); it not only enhances learning and retention of new information but also prevents the build-up of PI. Both variations in the encoding and

retrieval phases may contribute to the forward testing effect. Although interim testing may impair the learning of new information when tested and new materials are presented together, this negative effect can be eliminated and reversed by the separate presentation of tested and new information. Further investigations on aspects of this important effect, which are currently poorly understood, will additionally enhance our understanding.

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

Agarwal, P. K., & Roediger, H. L., 3rd. (2011). Expectancy of an open-book test decreases performance on a delayed closed-book test. *Memory, 19*(8), 836-852. doi: 10.1080/09658211.2011.613840

Anderson, J. R. (2000). *Learning and memory: An integrated approach (2nd ed.)*. Hoboken, NJ, US: John Wiley & Sons Inc.

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

Ashby, F. G., Alfonso-Reese, L. A., Turken, A. U., & Waldron, E. M. (1998). A neuropsychological theory of multiple systems in category learning. *Psychological Review, 105*(3), 442-481. doi: 10.1037/0033-295X.105.3.442

Aslan, A., & Bäuml, K. H. T. (2015). Testing enhances subsequent learning in older but not in younger elementary school children. *Developmental Science, 19*(6), 992-998. doi: 10.1111/desc.12340

Balzer, W. K., Doherty, M. E., & O'Connor, R. J. (1989). Effects of cognitive feedback on performance. *Psychological Bulletin, 106*(3), 410-433. doi: 10.1037/0033-2909.106.3.410

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

Bohay, M., Blakely, D. P., Tamplin, A. K., & Radvansky, G. A. (2011). Note taking, review, memory, and comprehension. *The American Journal of Psychology, 124*(1), 63–73. doi: 10.5406/amerjpsyc.124.1.0063

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2009). Converting among effect sizes. In: Introduction to meta-analysis. In U. Chichester (Ed.), *Introduction to meta-analysis* (pp. 45-49): John Wiley & Sons, Ltd.

Brady, T. F., Konkle, T., Alvarez, G. A., & Oliva, A. (2008). Visual long-term memory has a massive storage capacity for object details. *Proceedings of the National Academy of Sciences of the United States of America, 105*(38), 14325-14329. doi: 10.1073/pnas.0803390105

Bransford, J. D., & Johnson, M. K. (1972). Contextual prerequisites for understanding: Some investigations of comprehension and recall. *Journal of Verbal Learning and Verbal Behavior, 11*(6), 717-726. doi: 10.1016/S0022-5371(72)80006-9

Brooks, D. N. (1975). Long and short term memory in head injured patients. *Cortex, 11*(4), 329-340. doi: 10.1016/S0010-9452(75)80025-6

Butler, A. C. (2010). Repeated testing produces superior transfer of learning relative to repeated studying. *Journal of Experimental Psychology: Learning, Memory & Cognition, 36*(5), 1118-1133. doi: 10.1037/a0019902.

Butler, A. C., & Roediger, H. L. (2008). Feedback enhances the positive effects and reduces the negative effects of multiple-choice testing. *Memory & Cognition, 36*(3), 604-616. doi: 10.3758/mc.36.3.604

Button, K. S., Ioannidis, J. P. A., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S. J., & Munafò, M. R. (2013). Power failure: why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience, 14*, 365. doi: 10.1038/nrn3475

Callender, A. A., & McDaniel, M. A. (2009). The limited benefits of rereading educational texts. *Contemporary Educational Psychology, 34*(1), 30-41. doi: 10.1016/j.cedpsych.2008.07.001

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: 1037/a0017201

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

Chan, J. C., McDermott, K. B., & Roediger, H. L., 3rd. (2006). Retrieval-induced facilitation: initially nontested material can benefit from prior testing of related material. *Journal of Experimental Psychology: General, 135*(4), 553-571. doi: 10.1037/0096-3445.135.4.553

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, 70*(7), 1-60. doi: 10.1080/17470218.2016.1175485

Dai, X., & Gong, Y. (1987). A comparison of factor analytic studies among WAIS-RC, WAIS and WAIS-R. *Acta Psychologica Sinica, 1*, 70-78.

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

Davis, S. D., Chan, J. C. K., & Wilford, M. M. (2017). The dark side of interpolated testing: Frequent switching between retrieval and encoding impairs new learning. *Journal of Applied Research in Memory and Cognition, 6*(4), 434-441. doi: 10.1016/j.jarmac.2017.07.002

DeLozier, S., & Dunlosky, J. (2015). How do students improve their value-based learning with task experience? *Memory, 23*(6), 928-942. doi: 10.1080/09658211.2014.938083

Djonlagic, I., Rosenfeld, A., Shohamy, D., Myers, C., Gluck, M., & Stickgold, R. (2009). Sleep enhances category learning. *Learning & Memory, 16*(12), 751-755. doi: 10.1101/lm.1634509

Dunlosky, J., Rawson, K. A., Marsh, E. J., Nathan, M. J., & Willingham, D. T. (2013). Improving students' learning with effective learning techniques: Promising directions from cognitive and educational psychology. *Psychological Science in the Public Interest, 14*(1), 4-58. doi: 10.1177/1529100612453266

DuPaul, G. J., & Volpe, R. J. (2009). ADHD and learning disabilities: Research findings and clinical implications. *Current Attention Disorders Reports, 1*(4), 152. doi: 10.1007/s12618-009-0021-4

Eitel, A., & Kühl, T. (2015). Effects of disfluency and test expectancy on learning with text. *Metacognition and Learning, 11*(1), 107-121. doi: 10.1007/s11409-015-9145-3

Ellis, J. A., & Montague, W. E. (1973). Effect of recalling on proactive interference in short-term memory. *Journal of Experimental Psychology, 99*(3), 356-359. doi: 10.1037/h0035246

Emery, L., Hale, S., & Myerson, J. (2008). Age differences in proactive interference, working memory, and abstract reasoning. *Psychology and Aging, 23*(3), 634-645. doi: 10.1037/a0012577

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. doi: 10.3758/BF03193146

Fey, L. (2012, March). 'Drill and kill' testing: Just say "no". Retrieved from https://www.msdf.org/blog/2012/03/drill-and-kill-testing-just-say-no/

Finn, B. (2017). A framework of episodic updating: an account of memory updating after retrieval. *Psychology of Learning and Motivation, 67*, 173-211. doi: 10.1016/bs.plm.2017.03.006

Finn, B., & Roediger, H. L., III. (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

Gao, W., Cao, B., Shan, S., Chen, X., Zhou, D., Zhang, X., & Zhao, D. (2008). The CAS-PEAL large-scale Chinese face database and baseline evaluations. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans, 38*(1), 7-11.

Giguere, G., & Love, B. C. (2013). Limits in decision making arise from limits in memory retrieval. *Proceedings of the National Academy of Sciences of the United States of America, 110*(19), 7613-7618. doi: 10.1073/pnas.1219674110

GrÉGoire, J., & Van Der Linden, M. (1997). Effect of age on forward and backward digit spans. *Aging, Neuropsychology, and Cognition, 4*(2), 140-149. doi: 10.1080/13825589708256642

Harlen, W., & Deakin Crick, R. (2003). Testing and motivation for learning. *Assessment in Education: Principles, Policy & Practice, 10*(2), 169-207. doi: 10.1080/0969594032000121270

Hausman, H., & Kornell, N. (2014). Mixing topics while studying does not enhance learning. *Journal of Applied Research in Memory and Cognition, 3*(3), 153-160. doi: 10.1016/j.jarmac.2014.03.003

Healy, A. F., Jones, M., Lalchandani, L., & Tack, L. A. (in press). Timing of quizzes during learning: Effects on motivation and retention. *Journal of Experimental Psychology: Applied*. doi: 10.1037/xap0000123

Henry, J. D., MacLeod, M. S., Phillips, L. H., & Crawford, J. R. (2004). A meta-analytic review of prospective memory and aging. *Psychology and Aging, 19*(1), 27-39. doi: 10.1037/0882-7974.19.1.27

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

Hodges, J. R., Salmon, D. P., & Butters, N. (1992). Semantic memory impairment in Alzheimer's disease: Failure of access or degraded knowledge? *Neuropsychologia, 30*(4), 301-314. doi: https://doi.org/10.1016/0028-3932(92)90104-T

Holland, J. H., Holyoak, K. J., Nisbett, R. E., & Thagard, P. R. (1989). *Induction: Processes of inference, learning, and discovery.* MIT Press.

Ikier, S., & Hasher, L. (2006). Age differences in implicit interference. *The Journals of Gerontology Series B: Psychological Sciences and Social Sciences, 61*(5), 278-284. doi: 10.1093/geronb/61.5.P278

Ikier, S., Yang, L., & Hasher, L. (2008). Implicit proactive interference, age, and automatic versus controlled retrieval strategies. *Psychological Science, 19*(5), 456-461. doi: 10.1111/j.1467-9280.2008.02109.x

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

Jackson, J. D., & Balota, D. A. (2012). Mind-wandering in younger and older adults: converging evidence from the Sustained Attention to Response Task and reading for comprehension. *Psychology and Aging, 27*(1), 106-119. doi: 10.1037/a0023933

Jacoby, L. L., Wahlheim, C. N., & Coane, J. H. (2010). Test-enhanced learning of natural concepts: effects on recognition memory, classification, and metacognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 36*(6), 1441-1451. doi: 10.1037/a0020636

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

Jordano, M., L., & Touron, D., R. (2017). Stereotype threat as a trigger of mind-wandering in older adults. *Psychology and Aging, 32*(3), 307-313. doi: 10.1037/pag0000167.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

Karpicke, J. D., & Roediger, H. L., 3rd. (2008). The critical importance of retrieval for learning. *Science, 319*(5865), 966-968. doi: 10.1126/science.1152408

Knowlton, B. J., & Squire, L. R. (1993). The learning of categories: Parallel brain systems for item memory and category knowledge. *Science, 262*(5140), 1747-1749. doi: 10.1126/science.8259522

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. (2007). The promise and perils of self-regulated study. *Psychonomic Bulletin & Review, 14*(2), 219 -224. doi: 10.3758/BF03194055

Kornell, N., & Bjork, R. A. (2008a). Learning concepts and categories: Is spacing the "enemy of induction"? *Psychological Science, 19*(6), 585-592. doi: 10.1111/j.1467-9280.2008.02127.x

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

Kornell, N., Castel, A. D., Eich, T. S., & Bjork, R. A. (2010). Spacing as the friend of both memory and induction in young and older adults. *Psychology and Aging, 25*(2), 498-503. doi: 10.1037/a0017807

Kornell, N., Hays, M. J., & Bjork, R. A. (2009). Unsuccessful retrieval attempts enhance subsequent learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 35*(4), 989-998. doi: 10.1037/a0015729

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: 0.1037/0278-7393.32.3.609

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

Krawietz, S. A., Tamplin, A. K., & Radvansky, G. A. (2012). Aging and mind wandering during text comprehension. *Psychology & Aging, 27*(4), 951-958. doi: 10.1037/a0028831

Lee, H. S., & Ahn, D. (in press). Testing prepares students to learn better: The forward effect of testing in category learning. *Journal of Educational Psychology*. Advance online publication. doi: 10.1037/edu0000211

Leeming, F., C. (2002). Retrieving essential material at the end of lectures improves performance on statistics exams. *Teaching of Psychology, 38*(2), 94-97. doi: https://doi.org/10.1177/0098628311401587

Lehman, M., Smith, M. A., & Karpicke, J. D. (2014). Toward an episodic context account of retrieval-based learning: Dissociating retrieval practice and elaboration. *Journal of Experimental Psychology: Learning, Memory & Cognition, 40*(6), 1787-1794. doi: 10.1037/xlm0000012

Lin, L., & Fergus, I. M. C. (2008). Aging and memory: A cognitive approach. *The Canadian Journal of Psychiatry, 53*(6), 346-353. doi: 10.1177/070674370805300603

Love, B. C. (2000). Learning at different levels of abstraction. *Proceedings of the Cognitive Science Society, 22*, 800-805. Mahwah, NJ: Erlbaum.

Lustig, C., May, C. P., & Hasher, L. (2001). Working memory span and the role of proactive interference. *Journal of Experimental Psychology: General, 130*(2), 199-207. doi: 10.1037/0096-3445.130.2.199

Mathy, F., & Feldman, J. (2009). A rule-based presentation order facilitates category learning. *Psychonomic Bulletin & Review, 16*(6), 1050-1057. doi: 10.3758/PBR.16.6.1050

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

Middlebrooks, C. D., Murayama, k., & Castel, A. D. (in press). Test expectancy and memory for important information. *Journal of Experimental Psychology: Learning, Memory & Cognition*. Advance online publication. doi: 10.1037/xlm0000360

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

Murphy, G. L. (2002). *The big book of concepts*. Cambridge: MA: MIT Press.

Nelson, T. O., & Leonesio, R. J. (1988). Allocation of self-paced study time and the" labor-in-vain effect.". *Journal of Experimental Psychology: Learning, Memory, and Cognition, 14*(4), 676-686. doi: 10.1037/0278-7393.14.4.676

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

Nosofsky, R. M., Denton, S. E., Zaki, S. R., Murphy-Knudsen, A. F., & Unverzagt, F. W. (2012). Studies of implicit prototype extraction in patients with mild cognitive

impairment and early Alzheimer's disease. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 38*(4), 860-880. doi: 10.1037/a0028064

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

Old, S. R., & Naveh-Benjamin, M. (2008). Differential effects of age on item and associative measures of memory: A meta-analysis. *Psychology and Aging, 23*(1), 104-118. doi: 10.1037/0882-7974.23.1.104

Pashler, H., & Mozer, M. C. (2013). When does fading enhance perceptual category learning? *Journal of Experimental Psychology: Learning, Memory, and Cognition, 39*(4), 1162-1173. doi: 10.1037/a0031679

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

Pierce, B. H., Gallo, D. A., & McCain, J. L. (in press). Reduced interference from memory testing: A postretrieval monitoring account. *Journal of Experimental Psychology:*

Learning, Memory, and Cognition, Advance online publication. doi: 10.1037/xlm0000377

Piolino, P., Desgranges, B., Benali, K., & Eustache, F. (2002). Episodic and semantic remote autobiographical memory in ageing. *Memory, 10*(4), 239-257. doi: 10.1080/09658210143000353

Potts, R., Davies, G., & Shanks, D. R. (2018). The benefit of generating errors during learning: What is the locus of the effect? *Journal of Experimental Psychology: Learning, Memory & Cognition*, Advance online publication. doi: 10.1037/xlm0000637

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

Reed, J. M., Squire, L. R., Patalano, A. L., Smith, E. E., & Jonides, J. (1999). Learning about categories that are defined by object-like stimuli. *Behavioral Neuroscience, 113*(3), 411-419. doi: 10.1037/0735-7044.113.3.411

Roediger, H. L. III, Agarwal, P. K., McDaniel, M. A., & McDermott, K. B. (2011). Test-enhanced learning in the classroom: Long-term improvements from quizzing. *Journal of Experimental Psychology: Applied, 17*(4), 382-395. doi: 10.1037/a0026252

Roediger, H. L. III, & 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, 140*(6), 1432-1463. 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

Schwenn, E., & Postman, L. (1965). Studies of learning to learn: V. Gains in performance as a function of warm-up and associative practice. *Journal of Verbal Learning and Verbal Behavior, 6*(4), 565-573. doi: https://doi.org/10.1016/S0022-5371(67)80018-5

Stronge, J. H., Ward, T. J., & Grant, L. W. (2011). What makes good teachers good? A cross-case analysis of the connection between teacher effectiveness and student achievement. *Journal of Teacher Education, 62*(4), 339-355. doi: 10.1177/0022487111404241

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

DuPaul, G. J., & Volpe, R. J. (2009). ADHD and learning disabilities: Research findings and clinical implications. *Current Attention Disorders Reports, 1*(4), 152. doi: 10.1007/s12618-009-0021-4

Hodges, J. R., Salmon, D. P., & Butters, N. (1992). Semantic memory impairment in Alzheimer's disease: Failure of access or degraded knowledge? *Neuropsychologia, 30*(4), 301-314. doi: https://doi.org/10.1016/0028-3932(92)90104-T

Seli, P., Smallwood, J., Cheyne, J. A., & Smilek, D. (2015). On the relation of mind wandering and ADHD symptomatology. *Psychonomic Bulletin & Review, 22*(3), 629-636. doi: 10.3758/s13423-014-0793-0

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

Simpson, O. (2013). *Supporting students in online open and distance learning*. London: Routledge.

Smith, J. D., Redford, J. S., Washburn, D. A., & Taglialatela, L. A. (2005). Specific-token effects in screening tasks: possible implications for aviation security. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 31*(6), 1171-1185. doi: 10.1037/0278-7393.31.6.1171

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 of the United States of America, 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

Thomas, R. C., & McDaniel, M. A. (2013). Testing and feedback effects on front-end control over later retrieval. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 39*(2), 437-450. doi: 10.1037/a0028886

Thomaz, C. E., & Giraldi, G. A. (2010). A new ranking method for principal components analysis and its application to face image analysis. *Image and Vision Computing, 28*(6), 902-913. doi: 10.1016/j.imavis.2009.11.005

Thune, L. E. (1950). The effect of different types of preliminary activities on subsequent learning of paired-associate material. *Journal of Experimental Psychology, 40*(4), 423-438. doi: 10.1037/h0060560

Tse, C. S., Balota, D. A., & Roediger, H. L., 3rd. (2010). The benefits and costs of repeated testing on the learning of face-name pairs in healthy older adults. *Psychology and Aging, 25*(4), 833-845. doi: 10.1037/a0019933

Vadillo, M. A., Konstantinidis, E., & Shanks, D. R. (2016). Underpowered samples, false negatives, and unconscious learning. *Psychonomic Bulletin & Review, 23*(1), 87-102. doi: 10.3758/s13423-015-0892-6

Vanides, J., Yin, Y., Tomita, M., & Ruiz-Primo, M. A. (2005). Concept maps. *Science Scope, 28*(8), 27-31.

Wechsler, D. (1955). *Manual for the Wechsler Adult Intelligence Scale*. Oxford, England: Psychological Corp.

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

Whitebourne, S. K., & Whitebourne, S. B. (2014). *Adult development and aging: Biopsychosocial perspectives*: USA: John Willey & Sons.

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

Wnag, D., Shen, J., Peng, H., Tang, D., & Zhang, L. (2005). The model of educational effect on old adult's cognition. *Acta Psychologica Sinica, 37*(4), 511-516.

Yang, C., Potts, R., & Shanks, D. R. (2017). Metacognitive unawareness of the errorful generation benefit and its effects on self-regulated learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 43*(7), 1073-1092. doi: 10.1037/xlm0000363

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*(4), 991-1005. doi: 10.1037/edu0000031

Zaromb, F. M., & Roediger, H. L., 3rd. (2010). The testing effect in free recall is associated with enhanced organizational processes. *Memory & Cognition, 38*(8), 995-1008. doi: 10.3758/MC.38.8.995

Zhou, A., Yang, T., Cheng, C., Ma, X., & Zhao, J. (2015). Retrieval practice produces more learning in multiple-list tests with higher-order skills. *Acta Psychologica Sinica, 47*(7), 928. doi: 10.3724/sp.j.1041.2015.00928