The Forward Effects of Testing Transfer to Different Domains of Learning

Chunliang Yang [1], Siew-Jong Chew [1], Bukuan Sun [2], and David R. Shanks [1]

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

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

Author Note

All data have been made publicly available via the Open Science Framework (OSF) at https://osf.io/px274/. Correspondence concerning this article should be addressed to Chunliang Yang, Division of Psychology and Language Sciences, University College London, 26 Bedford Way, London WC1H 0AP. Email: chunliang.yang.14@ucl.ac.uk.

Author Contributions

Yang, Chew, and Shanks designed Experiment 1; Chew conducted Experiment 1; Yang and Chew analysed Experiment 1 data; Yang and Shanks designed Experiment 2; Sun conducted Experiment 2; Yang analysed Experiment 2 data; Yang and Shanks designed and conducted Experiment 3; Yang and Shanks analysed Experiment 3 data; all authors contributed to writing and approved the final version of this article.

Acknowledgments

An award from the China Scholarship Council (CSC) to Chunliang Yang supported this research. We thank five anonymous reviewers for their constructive comments.

Abstract

Interim testing of studied information, compared to restudying or no treatment, facilitates subsequent learning and retention of new information – *the forward testing effect*. Previous research exploring this effect has shown that interim testing of studied information from a given domain enhances subsequent learning and retention of new information within the same domain. In the current research, we ask whether interim testing can enhance subsequent encoding and retention of new information from a different domain. Experiment 1 showed that the forward testing effect is transferable; Experiment 2 further demonstrated this transferability even when material types and test formats are frequently switched; Experiment 3 documented transferability from low- to high-level learning. The results support a combined test-expectancy and retrieval-effort theory to account for the transfer of the forward testing effect.

Keywords: transferability; forward testing effect; interim testing; encoding

Educational Impact and Implications Statement

Given that people's study effort (e.g., attention, motivation) tends to decay across a study phase and attenuated study effort leads to a decline in learning efficiency and impairs learning outcomes, it is important to explore effective strategies to sustain study effort and maintain learning efficiency across a study phase. In the current research, we conceptually replicated the finding from previous experiments that interim testing of studied information (e.g., face-name pairs) facilitates the learning of new information in the same domain. Going beyond previous experiments, we show that testing of studied information from one domain (e.g., facts) also enhances the learning of new information from a different domain (e.g., visual concepts). These findings imply that administering interim low-stakes tests during a study phase can be profitably used to enhance the learning of new information, regardless of whether it is from the same or a different domain.

Although commonly considered as a measurement tool, testing has been repeatedly shown to be an efficient technique for improving learning and retention of both studied and new information (Pastötter & Bäuml, 2014; Roediger & Karpicke, 2006a; Yang, Potts, & Shanks, 2018). Over the last 100 years, researchers have demonstrated that testing is a more powerful strategy for consolidating and improving retention of studied information compared to restudying or doing nothing, even when no corrective feedback is provided in the tests – the *testing effect* (Abbott, 1909; Roediger & Karpicke, 2006b). We term the classic testing effect the *backward testing effect* in the current article following Pastötter and Bäuml (2014) and Yang, Potts, and Shanks (2017a). The backward testing effect has been shown to be a robust phenomenon across different educational materials (e.g., paired-associates, word lists, texts) and in different contexts (e.g., in laboratory research as well as in the classroom) (for reviews, see Roediger & Karpicke, 2006a; Rowland, 2014).

More recently, accumulating evidence has established that testing of studied information from one domain also facilitates learning and retention of new information within the same domain – the *forward testing effect*[1] (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; Yang et al., 2017a; Yang & Shanks, 2018). Szpunar et al. (2008) conducted a classic study demonstrating this facilitatory forward effect of testing. In their Experiment 2, five groups of participants were instructed to study five successive lists of words. The first group took an interim test (a free recall test) following each list. In contrast, the second group was only tested on List 2, the third group was only tested on List 3, the fourth group was only tested on List 4, and the fifth group was only tested on List 5. Szpunar et al.'s results demonstrated that the first group's correct recall was consistent across the five lists' interim tests, whereas the other four groups' correct recall systematically decreased as the number of previously untested lists increased. Meanwhile, proactive interference (PI; i.e., the interfering influence of words from prior lists on subsequent recall of the

[1] Many previous studies (e.g., Arnold & McDermott, 2013) found that testing of studied information enhances learning efficiency if the same information is restudied, a phenomenon commonly termed *test-potentiated learning*. In contrast to test-potentiated learning, the forward testing effect refers to the enhancing effect of testing on learning and retention of new information.

target list; the magnitude of PI is indexed by the number of incorrectly recalled words from prior lists)[2] did not significantly fluctuate across lists in the first group but substantially increased in the other groups as the number of previously untested lists increased. These results demonstrate that interim testing of studied lists enhances subsequent learning and recall of a new list.

The forward testing effect has been established as a robust phenomenon using a range of educational materials, including words (Aslan & Bäuml, 2015; Bäuml & Kliegl, 2013; Nunes & Weinstein, 2012; Pastötter et al., 2011; Pierce, Gallo, & McCain, in press; Weinstein et al., 2014; Yang et al., 2017a), line drawings of common objects (Pastötter, Weber, & Bäuml, 2013), foreign-translation pairs (Cho, Neely, Crocco, & Vitrano, 2016; Yang et al., 2017a), face-name pairs (Weinstein et al., 2011; Yang et al., 2017a), text passages (Healy, Jones, Lalchandani, & Tack, in press; Wissman, Rawson, & Pyc, 2011; Zhou, Yang, Cheng, Ma, & Zhao, 2015), lecture videos (Jing, Szpunar, & Schacter, 2016; Szpunar et al., 2013; Yue, Soderstrom, & Bjork, 2015), and paintings (Lee & Ahn, in press; Yang & Shanks, 2018) (for reviews, see Pastötter & Bäuml, 2014; Yang et al., 2018). Although the forward testing effect has been extensively researched, its underlying mechanisms are still unclear. Several theories have been proposed to account for this effect (for a more detailed discussion, see Yang et al., 2018). Below we briefly introduce those theories and explain the different predictions they make regarding transfer of the forward testing effect.

Context-change theory

Szpunar et al. (2008) suggested that interim testing induces context changes between lists, which improve list segregation and prevent the build-up of PI, and the release from PI produces greater recall of a new list. Interim testing of previously studied lists, compared with restudying or no treatment, modifies the mental contexts of these lists, and therefore these studied/tested lists are associated with both study and test contexts (Karpicke, Lehman, & Aue, 2014). By contrast, a new list

[2] The number of items incorrectly recalled from prior lists is a common but not the only measure of PI. For instance, in the modified Brown-Peterson paradigm (Wickens, Born, & Allen, 1963), participants study several lists of items from the same category (e.g., 3 lists of city names) and recall them shortly after studying each list. In this procedure the magnitude of PI is indexed by the decrease of correct recall (i.e., the number of items correctly recalled from the just-studied list) across lists.

studied subsequently is only associated with a study context. The difference in mental contexts induced by interim testing increases differentiation between studied/tested lists and a new list. Greater list differentiation improves list segregation, reduces the accumulation of PI, and facilitates recall of a new list.

Pastötter et al. (2011) proposed that interim testing induces mental context changes between lists, and those context changes create a "reset" of the encoding process and serve to maintain encoding engagement (e.g., attention) throughout a study phase. Pastötter et al.'s (2011) results support this explanation: Participants' attention decreased across lists when no interim tests were administered (indicated by an increase of alpha power - oscillatory brain activity inversely associated with attention), but this attenuation trend was reduced when interim tests were administered after each list.

In summary, both Szpunar et al. (2008) and Pastötter et al. (2011) proposed that context changes, induced by interim testing, play an important role in the forward testing effect: interim tests induce context changes between different study events, which reduce the build-up of PI and/or reset subsequent encoding. We term this explanation the context-change theory.

Strategy-change theory

Besides the context-change theory, Cho et al. (2016) proposed that the forward testing effect may be caused by encoding and/or retrieval strategy changes – a *strategy-change* theory. For example, while studying a few blocks of materials from the same domain (e.g., foreign-translation word pairs) and taking an interim test with the same format following studying each block (e.g., a cued recall test in which recall of translations in response to the foreign words is required), the prior interim tests may inform participants about the test format (e.g., cued recall). Accordingly, they may adapt their encoding strategies in order to encode new information more efficiently (for an illustration that prior interim tests induce encoding strategy changes, see Soderstrom & Bjork, 2014). More efficient encoding strategies optimize subsequent encoding and produce superior memory outcomes. Studying

the same type of material in a few successive blocks and taking the same type of interim test following studying each block may also induce more efficient retrieval strategies (for an illustration that prior interim tests induce retrieval strategy changes, see Thomas & McDaniel, 2013), which may in turn facilitate subsequent test performance.

Motivation theories: The context-change and strategy-change theories focus on the roles of non-motivational factors in the forward testing effect. Yang et al. (2018) proposed that the effect could be caused or mediated by changes in motivation: Prior interim tests may motivate learners to exert greater encoding and/or retrieval effort in the subsequent learning and test phases. Indeed, recent research has found that interim testing of studied information sustains learners' encoding engagement (e.g., study time, attention) across a study phase. For example, Szpunar et al. (2013) found that interim testing reduces mind-wandering while participants study a lecture video. More specifically, Jing et al. (2016) found that interim testing reduces task-irrelevant mind-wandering (off-task thoughts) but increases task-relevant mind-wandering (e.g., thoughts relating lecture content to real life). Jing et al. also demonstrated that task-relevant mind-wandering facilitates learning whereas task-irrelevant mind-wandering impairs it. Yang et al. (2017a) allowed participants to study a few lists of items at their own pace and study time allocation was measured as an index of learning effort. The results showed that in the absence of interim tests, participants' study time decreased across successive lists, whereas the decrease of study time was prevented by interim tests. Along the same lines, Healy et al. (in press) found that, while learning text statements, participants' self-ratings of learning engagement decreased significantly across a study phase in the absence of interim tests, but this decreasing trend was alleviated by interim testing. Collectively, these studies imply that interim testing of studied information sustains encoding engagement across a study phase, and maintained learning engagement may contribute to the forward testing effect. Two specific motivational theories have been proposed to account for why prior interim tests drive learners to sustain subsequent encoding engagement.

Test-expectancy theory: Weinstein et al. (2014) proposed that prior interim tests may act as warnings of a subsequent test, and induce greater test expectancy on the subsequent list, which in turn

motivates learners to exert more effort toward encoding the next list (Agarwal & Roediger, 2011). We term this theory the *test-expectancy* theory. Weinstein et al. (2014) explored the role of test expectancy in the forward testing effect. By asking participants to study 5 lists of words, Weinstein et al. again obtained a forward testing effect. More importantly, they found that their test group (who undertook a free recall test after studying each list) increased their test expectancy across lists, whereas their no-test group (which only undertook an interim test on the final list) decreased their test expectancy across the lists. A variety of studies have established that expecting a later test improves subsequent learning outcomes and test performance (e.g., Agarwal & Roediger, 2011; Eitel & Kühl, 2015; Middlebrooks, Murayama, & Castel, in press; Nestojko, Bui, Kornell, & Bjork, 2014).

Failure-encoding-effort theory: Cho et al. (2016) proposed a *failure-encoding-effort theory* to account for why prior interim tests enhance subsequent encoding effort. This theory suggests that retrieval failures (unsuccessful recall) in prior interim tests inform people about the difficulty of achieving successful recall and make them aware of the gap between their expected and actual learning levels, which then motivates them to exert more effort toward encoding new information to narrow the perceived gap. Put differently, this theory postulates that retrieval failures in prior interim tests induce dissatisfaction about the learning of prior information and this dissatisfaction motivates people to commit more effort to encode new information. Previous studies have shown that retrieval failures or committing errors in prior tests can potentiate encoding in a subsequent study phase (Kornell, Hays, & Bjork, 2009; Potts & Shanks, 2014; Yang, Potts, & Shanks, 2017b).

Besides enhanced encoding effort, greater retrieval effort may also contribute to the forward testing effect. For example, Cho et al. (2016) proposed a *retrieval-effort* theory to account for the forward testing effect. This theory hypothesizes that retrieval failures in prior interim tests induce dissatisfaction about recall performance in those tests, which motivates people to exert more effort to retrieve the new information in the subsequent interim test, and greater retrieval effort produces superior recall performance. To our knowledge, the key prediction of this theory has not been tested yet (see Experiment 3 for a detailed discussion).

In summary, five theories have been advanced to account for the forward testing effect. The context-changes theory focuses on the influence of context changes on subsequent learning and retrieval of new information. The strategy-change theory assumes encoding and retrieval strategy changes contribute importantly to the effect. Different from the context-change and strategy-change theories, three motivational theories (test-expectancy, failure-encoding-effort, and retrieval-effort) assume that the effect is caused or mediated by changes in motivation: Prior interim tests induce greater motivation to encode and retrieve new information.

Rationale for the current research

Many previous studies have documented that testing of studied information from one domain robustly enhances learning and retention of new information from the same domain – the *classic forward testing effect*. The five theories briefly described above all endeavor to explain this important finding. In those studies, the type of material has always been the same across lists/blocks. Whether or not the forward testing effect transfers across different domains of learning is unknown. The main aim of the current research 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*. The key difference between the current and previous studies is that we switched material types from prior to the target (final) blocks, which enabled us to test the transferability of the effect.

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 bears theoretical implications (see below for a detailed discussion).

Would we expect the forward testing effect to be transferable? Given that previous studies have shown that the effect is robust, intuitively we might anticipate an affirmative answer. At least some of the theories described above clearly predict transfer. For instance, 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. Another reason to anticipate transfer is that a recent meta-analysis of the backward testing effect, which combined 10,382 participants' data from 122 experiments, established that it is robustly transferable across different situations (e.g., from one test format to a different format, from tests on fact items to tests on inference questions, etc; for details, see Pan & Rickard, 2018).

However, there are at least four reasons to predict no or minimal transfer. First, a few of the aforementioned theories predict no or minimal transfer. For example, the context-change theory predicts little transfer of the effect because switching material types also induces substantial context changes between different learning events (e.g., Ellis & Montague, 1973; Emery, Hale, & Myerson, 2008; Lustig, May, & Hasher, 2001; Nunes & Weinstein, 2012), which will wipe out PI and "reset" subsequent encoding of new information regardless of whether interim tests are administered or not.

Second, even the test-expectancy and failure-encoding-effort 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 cannot 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.

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 minimal improvement in learning and recall of new information – the "*labor in vain effect*" (Callender & McDaniel, 2009; DeLozier & Dunlosky, 2015; Nelson & Leonesio, 1988; Yang & Shanks, 2018) – and lead to little transferability of the forward testing effect.

Fourth, previous *transfer of training* research either failed to achieve successful transfer of training or found that the knowledge or skills obtained from the trained tasks only benefits the performance in the trained or similar tasks but fail to generalise to other different tasks – *near transfer of training* (for a review, see Grossman & Salas, 2011). These findings imply that the forward testing effect may transfer minimally to different domains of learning.

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 research explores this important issue – the transferability of the forward testing effect.

Experiment 1

Experiment 1 had three aims. The first was to conceptually replicate the classic forward testing effect (i.e., testing of studied information in one domain enhances learning and retention of new information in the *same* domain). To achieve this, we employed two groups (Same-Test and Same-Math) of participants 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 one domain enhances learning and retention of new information in a *different* domain). To achieve this, we employed two other groups (Different-Test and Different-Math) of participants who studied 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 across lists in the test group but decreases in the no-test group), and therefore all participants were asked to report their test expectancy before studying each list.

Method

Participants

Given that no previous studies have explored the transfer of the forward testing effect, we therefore determined the sample size according to previous non-transfer forward testing effect studies (Weinstein et al., 2011; Yang et al., 2017a), in which the observed effect sizes (Cohen's *d*s) of the forward testing effect in the learning of face-name pairs ranged from 0.87 to 1.47. Using these effect sizes and the G*Power program (Faul, Erdfelder, Lang, & Buchner, 2007), we calculated that about 8-29 participants in each group are required to observe a significant (α = .05; power = 0.90) forward testing effect. We therefore included 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.[3] No participants had previously taken part in Yang et al.'s (2017a) or Yang and Shanks's (2018) 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 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

[3] The current research was an international collaboration project: Experiments 1 and 3 were conducted in the UK and Experiment 2 was run in China. Experiment 3 was conducted in a UCL psychology class as part of a class requirement, enabling us to collect a large set of data in one session.

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 studied three lists of Swahili-English pairs and then a list of face-name pairs. All four groups were warned at the outset about the final 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. 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 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 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 questions blank if they did not remember the answers.

Results

Scoring

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

List 1-3 interim test recall

Table 1 reports 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 significantly fluctuate across lists, $F(2,78) = 2.59$, $p = .08$, $\eta_p^2 = .06$, and there was no interaction between Material and List, $F(2,78) = .35$, $p = .70$, $\eta_p^2 = .01$. 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 1, the face-name pairs were more difficult to remember than the Swahili-English pairs.

List 4 interim test recall

Figure 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, $F(1,78) = 13.95$, $p < .001$, $\eta_p^2 = .15$, indicating that interim testing, compared to no interim testing (solving math problems), enhanced learning and retention of new information. There was a main effect of Material, $F(1,78) = 6.28$, $p = .01$, $\eta_p^2 = .08$, reflecting the fact that a switch of material type enhanced recall. There was no interaction between Interim task and Material, $F(1,78) < 0.001$, $p = .99$, $\eta_p^2 < .001$.

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 = 1.70 names, 95% CI = [0.31, 3.09], $t(38) = 2.48$, $p = .018$, Cohen's $d = 0.70$, indicating that interim testing of studied

information from one domain enhances learning and retention 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], $t(40) = 2.82$, $p = .007$, Cohen's $d = 0.87$, revealing that interim testing of studied information from one domain enhances learning and retention of new information from a *different* domain, and that the forward testing effect is to some degree transferable. An independent-samples t test indicated 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], $t(38) = -3.86$, $p < .001$, Cohen's $d = -1.22$, indicating that interim testing prevents the build-up of PI. This result suggests that the classic forward testing effect (i.e., the difference in List 4 interim test recall between the Same-Test and Same-Math groups) can partially be attributed to the context-change mechanism.

The amounts of PI in the Same-Test and Same-Math groups were significantly greater than 0, $t(19)$'s > 5.71, p's < .001, Cohen's d's > 1.27. In contrast, neither the Different-Test nor the Different-Math groups experienced any PI (i.e., there were no trials in which a participant recalled a Swahili word's translation from Lists 1-3 on the List 4 face-name test). These results imply that the finding that a switch of material type 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 types 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 context-change theory as an explanation for the transferability of the forward testing effect: PI was equivalent in the Different-Test and Different-Math groups, yet the former outperformed the latter in the critical List 4 test. We discuss theoretical implications more fully in the General 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 amount of current list intrusions, replicates previous findings (Weinstein et al., 2011; Yang et al., 2017a), indicating that the Same-Test group's memory search set in the List 4 interim test was smaller and providing some support for the context-change theory to account for the classic forward testing effect.

Cumulative test recall

Figure 2B depicts cumulative test recall of 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 factors: (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) a backward testing effect (i.e., testing of studied information enhances retention of that information compared to solving math problem); (3) a 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 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, $F(1,78) = 1.52$, $p = .22$, $\eta_p^2 = .02$.

Test expectancy ratings

Figure 2D depicts all four groups' test expectancy ratings across lists. The Same-Test and Different-Test groups gradually increased their test expectancy but the Same-Math and Different-Math groups gradually decreased their expectancy across lists. Because the test expectancy ratings were noisy, we collapsed the data across groups to increase the power to observe possible effects: the Same-Test and Different-Test groups were collapsed as a Test group, and 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, found 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, $F(1, 40) = 1.78$, $p = .19$, $\eta_p^2 = .04$[4], 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], $t(80) = 3.23$, $p = .002$, Cohen's $d = 0.71$, but there was no significant difference on any of Lists 1-3, $t(80)$'s < 0.82, p's $> .42$, Cohen's d's < 0.18. These results are consistent with Weinstein et al.'s (2014) findings and provide some support for the key process assumed to be critical according to the test-expectancy theory. We will return to the relationship between test expectancy and interim test recall below.

Discussion

The Same-Test group outperformed the Same-Math group in the List 4 interim test, replicating the classic forward testing effect. More novel is the finding that 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.

[4] The quadratic trend of the Test group's test expectancy across lists was significant, $F(1, 40) = 9.77$, $p = .003$, $\eta_p^2 = .20$.

Experiment 1 showed no interaction between Interim task (test/math) and Material (same/different) in List 4 interim test recall, indicating that a switch of material types did not significantly moderate the forward testing effect. However, as observed in Experiment 1 (see Table 1 and Figure 2B), the Swahili-English word pairs were easier to remember than the face-name pairs, which might have contributed to the non-significant interaction. Because of the difference in difficulty between Swahili-English word pairs and face-name pairs, it is premature to make a firm conclusion about whether switching of material types attenuates the forward testing effect. Given that the main aim of the current research is primarily to ask whether the forward testing effect can transfer, we did not address this question in the following two experiments. Future research could profitably explore this.

Experiment 2

In Experiment 2, we omitted the Same-Test, Same-Math, and Different-Math groups. A Different-Restudy group, which restudied the prior block's materials and was tested on the final (target) block, was added, which enabled us to explore the transferability of the forward testing effect by comparing interim testing with restudying.

Unlike Experiment 1, in Experiment 2 corrective feedback was offered in all interim tests. There is both a theoretical and a practical rationale for this change. Providing corrective feedback equates the Different-Test and Different-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. Providing 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 2 employed a more naturalistic procedure.

In Experiment 1, 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 transfer obtained in Experiment 1 might result from encoding- and retrieval-strategy changes that support performance in the final target list.

The primary goal of Experiment 2 was to explore the transferability of the forward testing effect when both material types and test formats are switched. Such switching will minimize any beneficial contribution from the strategy-change mechanisms proposed by Cho et al. (2016), because there is little reason to assume that effective encoding/retrieval strategies, developed and adopted across prior blocks, would be applicable in the final target block with a different type of material and test format. To achieve this, in Experiment 2, material types were changed from block to block: Block 1: object pictures; Block 2: text passage; Block 3: face-profession pairs. In addition, interim test formats were also changed from block to block: Block 1: recognition test; Block 2: fill-in-the-blank test; Block 3: cued recall test.

Test expectancy ratings in Experiment 1 were relatively noisy, which might have arisen from the fact that the rating scale (0-100) was too granular. In Experiment 2, we narrowed the rating scale (1-7). In Experiment 1, 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 2, we instead asked participants 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 1, the effect size of the forward testing effect was $d = 0.84$ in the transfer (Different) groups. The calculated sample size to observe a significant ($\alpha = .05$; power = .90) forward

testing effect in Experiment 2 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, China. 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). 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 interim restudy phases; the first and third sets were used in Block 1's cumulative test phase.

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 length. The first paragraph described the properties of graphene, the second concerned its uses, and the third was mainly about the history of research on 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 final cumulative test. They were also informed that the computer would randomly decide the next 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 the materials from Blocks 1 and 2 and undertook an interim test on Block 3 (see Figure 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 group viewed 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, and participants were informed that they only needed to remember the old pictures.

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 completed 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 under no time pressure. 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.

Results

Scoring

The computer automatically scored the test performance in all interim and cumulative tests.

Interim test performance of 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 false alarm (i.e., mistakenly judging new pictures as old) rate was 22.8% (*SD* = 14.37). Their 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 13.28/30 (*SD* = 10.01) target items.

Block 3 interim test recall

Of critical interest is the Block 3 interim test recall in the two groups. An independent-samples *t* test showed that the Different-Test group (*M* = 8.50, *SD* = 6.54) correctly recalled about twice many professions as the Different-Restudy group (*M* = 4.89, *SD* = 4.17), difference = 3.62 professions, $t(64) = 2.70$, $p = .009$, 95% CI = [0.94, 6.30], Cohen's $d = 0.66$ (see Figure 4A), again revealing that the forward testing effect transfers robustly.

Cumulative test performance

For Block 1 items in the cumulative test, as can be seen in Figure 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], $t(31) = 13.77$, $p < .001$, Cohen's d = 2.42; for the Different-Restudy group: d' = 1.51, 95% CI = [1.13, 1.89], $t(33) = 8.09$, $p < .001$, Cohen's d = 1.39. 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], $t(64) = 2.38$, $p = .021$, Cohen's d = 0.59. This result indicates a clear backward testing effect on recognition memory (Jacoby, Wahlheim, & Coane, 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], $t(64) = 3.93$, $p < .001$, Cohen's d = 0.97 (see Figure 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], $t(64) = 1.09$, $p = .28$, Cohen's d = 0.27 (see Figure 4D), the same qualitative pattern as observed in the Block 3 interim test. Previous studies showed that testing potentiates subsequent encoding of corrective feedback (Butler & Roediger, 2008; Potts & Shanks, 2014; Yang et al., 2017b). As can be seen in Figures 4A and 4D, the Different-Restudy group benefited much more from the Block 3 interim test than the Different-Test group: Recall improvement from the Block 3 interim test to the Block 3 cumulative test was smaller in the Different-Test group (recall improved from 8.50 to 9.84 items) than in the Different-Restudy group (from 4.89 to 8.24). 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 4E. A mixed ANOVA, with Block (1-3) as a within-subjects variable and Interim task as a between-subjects variable, showed that the Different-Test group developed higher test expectancy 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, $F(1, 31) = 12.07$, $p = .002$, $\eta_p^2 = .28$, [5] 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], $t(64) = 0.03$, $p = .98$, Cohen's $d = 0.006$, but the differences were significant in Block 2, difference = 1.02, 95% CI = [0.42, 1.62], $t(64) = 3.38$, $p = .001$, Cohen's $d = 0.84$, and Block 3, difference = 1.85, 95% CI = [1.04, 2.65], $t(64) = 4.59$, $p < .001$, Cohen's $d = 1.44$. We discuss the relationship between test expectancy and interim test recall more fully below.

Discussion

Experiment 2 again revealed that the forward testing effect is transferable. Going beyond Experiment 1, Experiment 2 demonstrates substantial transfer even when both material types and test formats are changed from block to block. Test expectancy ratings again are similar to those obtained by Weinstein et al. (2014). In addition, Experiment 2 replicated the transfer effect using a culturally different sample.

Experiment 3

Experiments 1 and 2 demonstrated that the forward testing effect transfers across different domains of relatively low-level learning (i.e., remembering specific items). The main aim of Experiment 3 is to explore whether the effect transfers from low- to high-level learning (e.g., inductive learning). Different from item learning, inductive learning is a process where learners are

[5] The quadratic trend of the Different-Test group's test expectancy across blocks was also significant, $F(1, 31) = 5.18$, $p = .03$, $\eta_p^2 = .14$.

required to abstract rules from a set of exemplars; Different from tests on specific items, induction tests require generalization of previous experience when making uncertain inferences that go beyond direct experience (for a detailed discussion about the differences between low- and high-level learning, see Yang & Shanks, 2018). The second aim of Experiment 3 is to test the transferability of the forward testing effect using more educationally-realistic materials.

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 3 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 3 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, we measured participants' response times (RTs) in the test stage of the final (target) block. RTs were taken as an index of retrieval effort (Pyc & Rawson, 2009). 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. Put differently, because of retrieval failures in prior interim tests, the Different-Test group may respond to test questions more cautiously and conservatively in the target block test, and hence longer RTs should be observed.

Experiment 2 demonstrated that interim testing with corrective feedback enhances subsequent learning more effectively than restudying. In Experiment 3, we omitted corrective feedback in the interim tests, which allowed us to directly compare the effect of interim testing with that of restudying on subsequent learning and retention of new information, removing any influences from additional learning via corrective feedback.

Unlike in Experiments 1 and 2, we also omitted the final cumulative test in Experiment 3. The main interest of Experiment 3 is the Block 4 interim test performance, and we had 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 our Experiments 1 and 2 – 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; Yang et al., 2017a; Yang & Shanks, 2018).

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 2 (Cohen's $d = 0.66$), the power to observe a significant ($\alpha = .05$) forward testing effect in Experiment 3 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 relatively unfamiliar artists (e.g., Philip Juras, Ryan Lewis). The paintings were taken from Kornell and Bjork (2008). 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 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). Participants were instructed to imagine themselves as art students 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 either 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 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 they studied 10 new statements.

In the Block 4 study phase, the 48 paintings were shown one by one, for 5 sec each, with the artist's name given below. The paintings were presented in a spaced arrangement, following Kornell and Bjork (2008). 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 shown one by one in a random order, with the 8 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.

Results

Scoring

In the Block 1-3 tests, we assigned 2 points to correct answers and 1 point to partially correct answers. Yang scored the Block 1-3 test responses for the Different Test group. [6] 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], $t(130) = 4.44$, $p < .001$, Cohen's $d = 0.66$ (see Figure 6A), revealing that the forward testing effect is robustly transferable from verbal fact to visual concept learning. [7]

Retrieval effort in the Block 4 test

[6] 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.

[7] 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 transfer.

We calculated the mean RT in the Block 4 test for each participant (see Figure 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], $t(130) = 2.28$, $p = .024$, Cohen's $d = 0.40$, consistent with the retrieval-effort theory. [8] Further analyses explored whether classification accuracy (correct/incorrect) moderated the effect of prior interim tests (i.e., Block 1-3 tests) on subsequent retrieval effort (i.e., RTs in the Block 4 test). Appendix A reports the detailed results.

Test expectancy ratings

Mean test expectancy ratings, shown in Figure 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, $F(1, 67) = 22.25$, $p < .001$, $\eta_p^2 = .25$. [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], $t(130) = -0.58$, $p = .57$, Cohen's $d = -0.10$, and 2, difference = 0.35, 95% CI = [-0.49, 1.18], $t(130) = 0.82$, $p = .42$, Cohen's $d = 0.14$. However, there were significant differences in Blocks 3, difference = 2.21, 95% CI = [1.35, 3.01], $t(130) = 5.12$, $p < .001$, Cohen's $d = 0.89$, and 4, difference = 3.06, 95% CI = [2.33, 3.80], $t(130) = 5.12$, $p < .001$, Cohen's $d = 0.89$. We will return to the relationship between test expectancy and test performance below.

[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], $t(130) = 2.08$, $p = .037$, Cohen's $d = 0.36$.

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

Discussion

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 showed the same pattern as Weinstein et al.'s (2014) findings.

Relationship between test expectancy and test recall

All three experiments consistently showed that the test groups (i.e., the Same-Test group in Experiment 1 and the Different-Test groups in Experiments 1-3) reported higher test expectancy than the control groups (i.e., the Same-Math and the Different-Math groups in Experiment 1 and the Different-Restudy groups in Experiments 2 and 3) on the target list/block. To determine directly whether test expectancy contributes to the forward testing effect, we conducted correlation analyses on the relationship between test expectancy and interim test recall (the following analyses were not pre-planned).

We collapsed the List 4 test expectancy ratings and test recall across the Same-Test and Same-Math groups in Experiment 1. These were significantly correlated, $r_{(40)} = .35$, $p = .03$, revealing an association between test expectancy and recall, consistent with the idea that expectancy contributes to the same-materials forward testing effect.

We then asked whether test expectancy contributes to the transfer of the forward testing effect. We collapsed the List 4 expectancy ratings and test recall data across the Different-Test and Different-Math groups in Experiment 1. These were not significantly correlated, $r_{(40)} = .11$, $p = .47$. Similarly, we collapsed the Block 3 expectancy ratings and recall data across the Different-Test and Different-Restudy groups in Experiment 2: $r_{(66)} = .18$, $p = .16$. Lastly, there was also no significant correlation between Block 4 expectancy ratings and test performance in Experiment 3, $r_{(132)} = .13$, $p = .13$. Figure 7 depicts the associations between test expectancy and recall in Experiments 1-3.

These results imply no significant correlation between test expectancy and recall, challenging the proposal that test expectancy contributes to the transfer 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, we consistently found a positive (although non-significant) correlation between test expectancy ratings and recall. Therefore, we conducted a meta-analysis to increase power.

In the meta-analysis, we excluded the Same-Test and Same-Math groups from Experiment 1 because the main aim of this meta-analysis is to explore whether test expectancy contributes to the transfer of the forward testing effect.[10] Using formulae explained by Borenstein, Hedges, Higgins, and Rothstein (2009), we first transformed the *r* values into Cohen's *d*s using the formula:

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

We calculated the variances of the *r* values using the formula:

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

where *N* represents the sample size. We next calculated the variances of Cohen's *d*s using the formula:

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

We then inserted these Cohen's *d*s and V_ds into the R *metafor* package and conducted a random effects meta-analysis. This revealed a significant albeit modest effect of test expectancy on recall, Cohen's d = 0.28, 95% CI = [0.03, .54] (see Figure 8 for detailed results).[11] Finally, we transformed this effect size (Cohen's *d*) back to *r* using the formula:

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

[10] Including Experiment 1's Same-Test and Same-Math groups does not change the pattern of results.

[11] Given that Experiment 1's test expectancy ratings were relatively noisy, we conducted a new random effects meta-analysis, in which the Different-Test and Different-Math groups from Experiment 1 were excluded. The new meta-analysis also showed a significant effect of test expectancy on interim test recall, Cohen's d = 0.30, 95 % CI = [0.01, 0.58]. The transformed *r* value is .15.

This yielded an *r* value of .14, confirming a weak correlation between test expectancy and recall.[12]

In summary, the significant correlation between test expectancy and recall performance in the Same groups (Same-Test and Same-Math) in Experiment 1 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 recall when the material is changed, supporting the test-expectancy theory as an account for the transfer of the forward testing effect.

General 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; Yang, Potts, & Shanks, 2017a; Yang & Shanks, 2017). The current research 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 1, 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, both the Different-Test and Different-Math groups in the List 4 interim test did not experience any PI.

Experiment 2 again confirmed this transfer. More importantly, it revealed substantial transfer even when both material type and test format are changed from block to block. Experiment 3 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 1-3) but decreased in the

[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). Using this method, we also observed a significant correlation between test expectancy and test performance: $r = 0.14$, $p = .024$.

control groups (the Same-Math and Different-Math groups in Experiment 1 and the Different-Restudy groups in Experiments 2 and 3). These test expectancy ratings were consistent with 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 transfer of the forward testing effect (i.e., the forward testing effects observed in Experiments 1-3's Different groups). As discussed above, the context-change mechanism should contribute little to transfer because switching material types also induces substantial context changes regardless of whether interim tests are administered or not. Moreover, although the strategy-change mechanisms might have contributed to the transfer findings in Experiment 1 – 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 strategy-change theory has difficulty explaining the transfer findings in Experiments 2 and 3, in which material types and test formats were both switched.

In contrast, the motivation theories readily explain the transfer findings in Experiments 2 and 3. The current research evaluated two motivation theories: test-expectancy and retrieval-effort. In all three experiments, we consistently observed 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. Besides test-expectancy, retrieval-effort also appears to contribute to the transfer we observed. As shown in Experiment 3, prior interim tests induced participants to spend longer responding to test questions. To our knowledge, Experiment 3 is the first to directly test the retrieval-effort theory.

Although Experiments 1 and 2 did not directly measure retrieval effort in the target block interim tests, retrieval effort might also have contributed to the transfer findings observed in these two experiments. Overall, the findings obtained in the current research support the test-expectancy and

retrieval-effort theories as a combined account for the forward testing effect. It must be noted that, although our results support these two theories, the current research does not exclude other (e.g., context-change, strategy-change) theories because they are not mutually exclusive. As noted by Yang et al. (2018, p. 6), different mechanisms may contribute to the forward testing effect in different situations, and in some situations different mechanism may operate in parallel producing overlapping forward testing effects. Further investigation on the underlying mechanisms is needed.

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 research, 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.

Practical (educational) implications

Learners' study effort (e.g., attention) and learning effectiveness tend to decay across a study phase. 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 1 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 1 demonstrated transfer; Experiment 2 revealed that the forward testing effect transfers even when material types and test formats are changed from block to block; and Experiment 3 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 research 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. Experiment 1 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.

Meta-analysis of a large body of research has established transferability of the backward testing effect (Pan & Rickard, 2018), and the current research demonstrates transfer of the forward testing effect. These findings jointly demonstrate the generalisability of the benefits of testing and encourage learners and instructors to use tests as a practical technique to improve learning and teaching outcomes.

Limitations

Given that the main aim of the current research was to explore transfer of the forward testing effect, substantial changes were induced across different blocks (e.g., material types and test formats) and across different experiments (e.g., feedback in interim tests was provided in Experiment 2 but not in Experiments 1 and 3). The forward testing effect survived even though these substantial changes were implemented, which testifies the robustness of the transfer. However, the changes across experiments might obfuscate the potential roles of certain variables. For example, corrective feedback may play a role in the transfer of the forward testing effect. According to the failure-encoding-effort theory, corrective feedback may inform participants about the number of retrieval failures in prior interim tests and induce dissatisfaction about prior learning, which in turn would motivate them to

devote more effort toward learning new information. Future research is needed to explore whether corrective feedback moderates the forward testing effect.

Because the observed correlation between test expectancy and recall performance was modest, the sample sizes in the current research were inadequate to yield a firm conclusion about the casual relationships amongst interim tests, test expectancy, and test performance. Future research could profitably explore this through increasing sample size and conducting mediation analyses.

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.

References

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

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

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*, 940-945. doi: 10.1037/a0029199

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

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

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*, 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*, 717-726. doi: 10.1016/S0022-5371(72)80006-9

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

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

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*, 1211-1235. doi: 10.1080/17470218.2016.1175485

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

Eitel, A., & Kühl, T. (2015). Effects of disfluency and test expectancy on learning with text. *Metacognition and Learning, 11*, 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*, 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*, 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*, 175-191. doi: 10.3758/BF03193146

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*, 149-161. doi: 10.1109/TSMCA.2007.909557.

Grossman, R., & Salas, E. (2011). The transfer of training: What really matters. *International Journal of Training and Development, 15*, 103-120. doi: 10.1111/j.1468-2419.2011.00373.x

Hausman, H., & Kornell, N. (2014). Mixing topics while studying does not enhance learning. *Journal of Applied Research in Memory and Cognition, 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

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

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

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

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

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*. doi: 10.1037/edu0000211

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

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

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*, 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*, 1038-1048. doi: 10.3758/s13421-014-0416-z

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

Pan, S. C., & Rickard, T. C. (2018). Transfer of test-enhanced learning: Meta-analytic review and synthesis. *Psychological Bulletin, 144*(7), 710-756. doi: 10.1037/bul0000151

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., Schicker, S., Niedernhuber, J., & Bäuml, K. H. (2011). Retrieval during learning facilitates subsequent memory encoding. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 37*, 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*, 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*. doi: 10.1037/xlm0000377

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

Pyc, M. A., & Rawson, K. A. (2009). Testing the retrieval effort hypothesis: Does greater difficulty correctly recalling information lead to higher levels of memory? *Journal of Memory and Language, 60*, 437-447. doi: 10.1016/j.jml.2009.01.004

Roediger, H. L., & Karpicke, J. D. (2006a). The power of testing memory: Basic research and implications for educational practice. *Perspectives on Psychological Science, 17*, 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*, 249-255. doi: 10.1111/j.1467-9280.2006.01693.x

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

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

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

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

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

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,* 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,* 902-913. doi: 10.1016/j.imavis.2009.11.005

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

Weinstein, Y., Gilmore, A. W., Szpunar, K. K., & McDermott, K. B. (2014). The role of test expectancy in the build-up of proactive interference in long-term memory. *Journal of*

Experimental Psychology: Learning, Memory, and Cognition, 40, 1039-1048. doi: 10.1037/a0036164.supp

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

Wickens, D. D., Born, D. G., & Allen, C. K. (1963). Proactive inhibition and item similarity in short-term memory. *Journal of verbal learning and verbal behavior, 2*(5), 440-445. doi: https://doi.org/10.1016/S0022-5371(63)80045-6

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

Yang, C., Potts, R., & Shanks, D. R. (2017a). The forward testing effect on self-regulated study time allocation and metamemory monitoring. *Journal of Experimental Psychology: Applied, 23*(3), 263-277. doi: 10.1037/xap0000122

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

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

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

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

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*, 928. doi: 10.3724/sp.j.1041.2015.00928

Table 1. *M* (*SD*) of List 1-3 interim test recall in Experiment 1.

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)

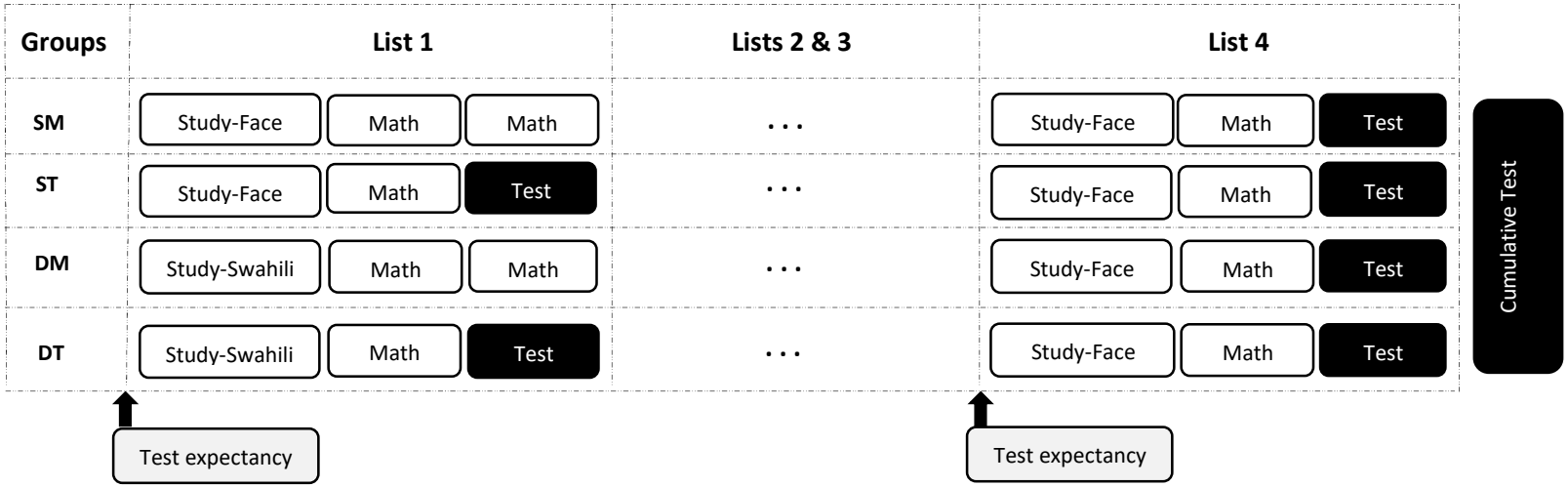

Figure 1. Experiment 1. The Same-Test (ST) and Same-Math (SM) groups studied four lists of face-name pairs while the Different-Test (DT) and Different-Math (DM) groups studied three lists of Swahili-English pairs followed by a list of face-name pairs. Prior to studying each list, all four groups reported 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.

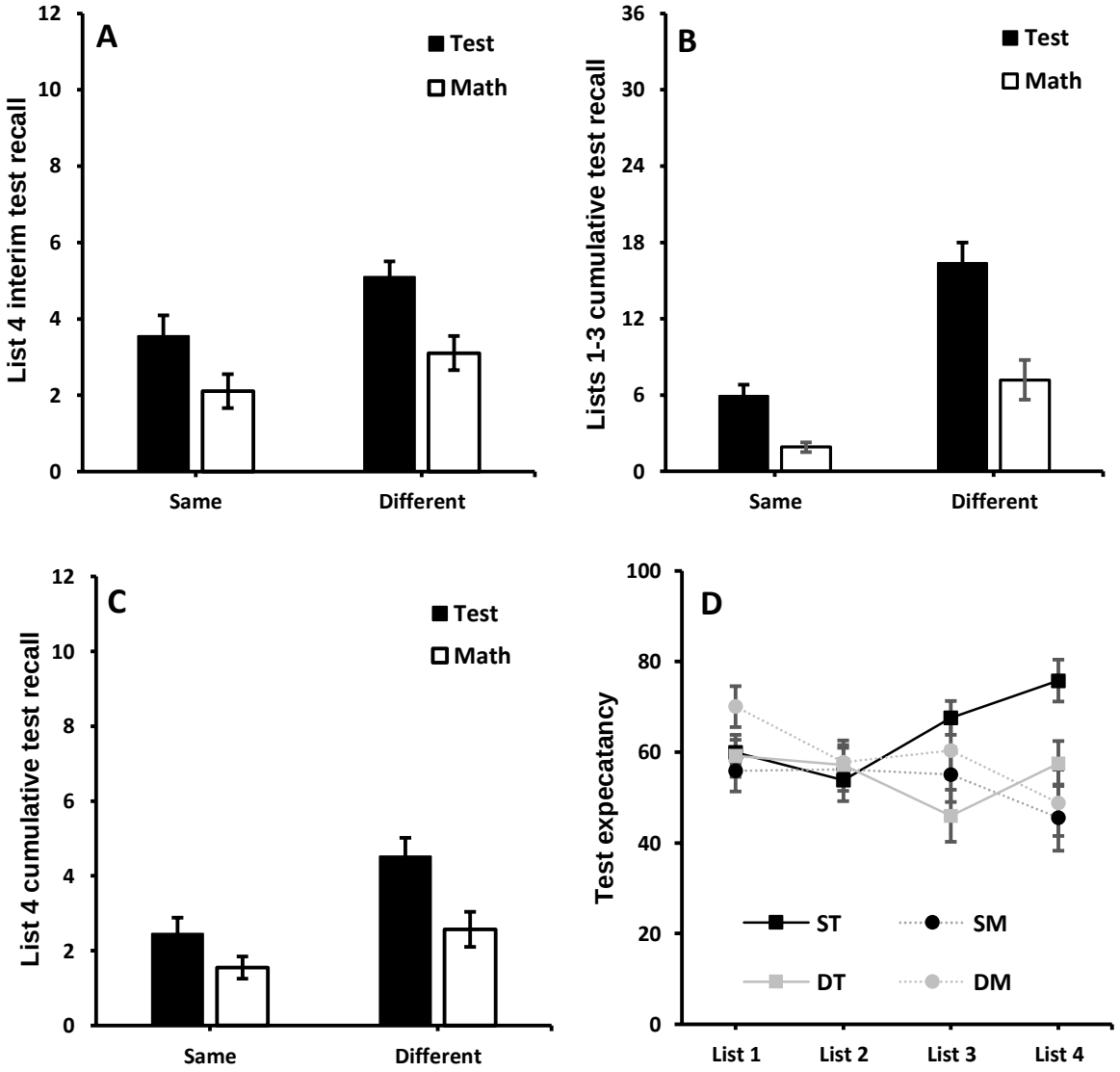

Figure 2. Experiment 1. Panel A: List 4 interim test recall; Panel B: Cumulative test recall of List 1-3 items; Panel C: Cumulative test recall of List 4 items; Panel D: Test expectancy ratings. ST = Same-Test; SM = Same-Math; DT = Different-Test; DM = Different-Math. Error bars represent ± 1 standard error.

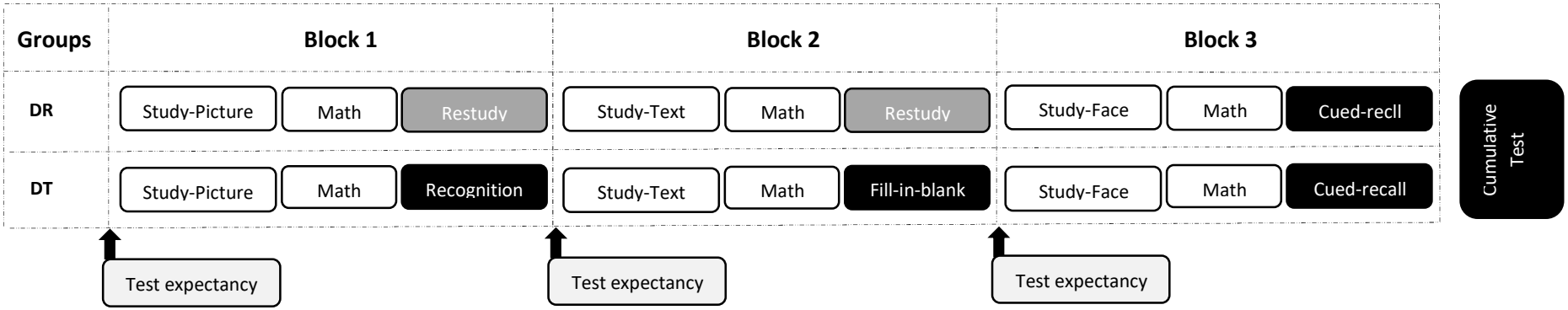

Figure 3. Experiment 2. The Different-Test (DT) and Different-Restudy (DR) groups studied different types of material across three blocks: Block 1: object pictures; Block 2: text; Block 3: face-profession pairs. Prior to studying each block, both groups reported their test expectancy. The Different-Test 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.

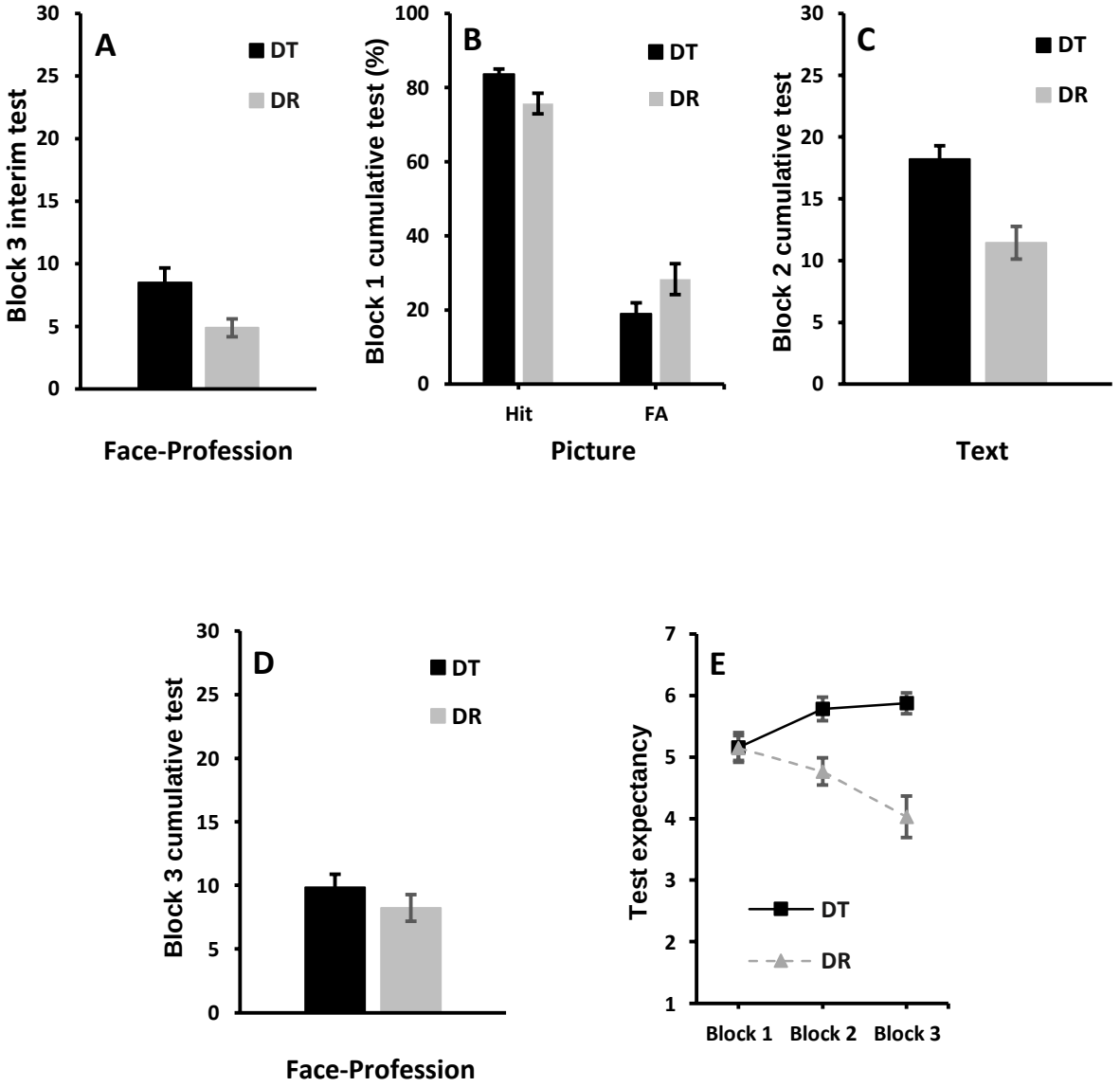

Figure 4. Experiment 2. Panel A: Block 3 interim test recall; Panel B: Hit and false alarm (FA) rates in the cumulative test for Block 1 items; Panel C: Cumulative test recall for Block 2 items; Panel D: Cumulative test recall for Block 3 items; Panel E: Test expectancy ratings. DT = Different-Test; DR = Different-Restudy. Error bars represent ± 1 standard error.

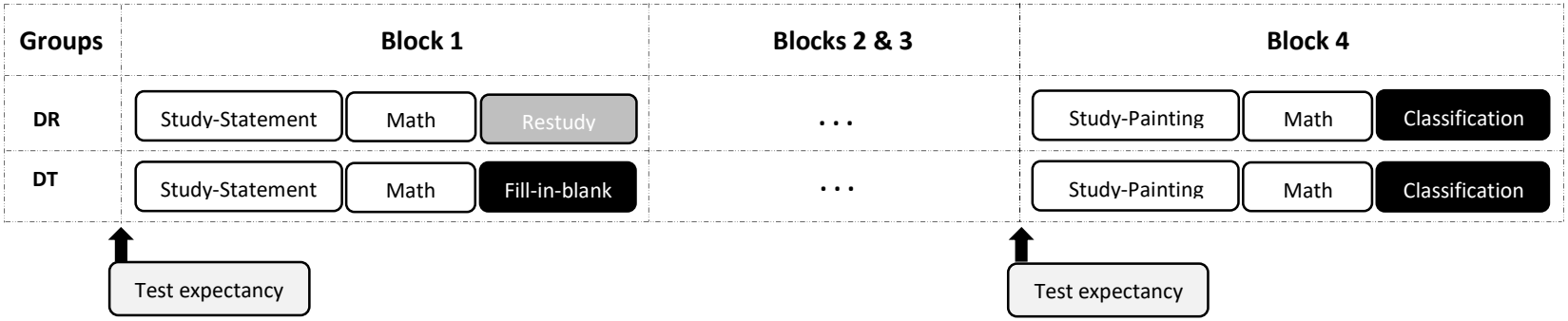

Figure 5. Experiment 3. The Different-Test (DT) and Different-Restudy (DR) groups studied three blocks of statements followed by a block of paintings. Prior to studying each block, both groups reported their test expectancy. The Different-Test group took tests on all four blocks whereas the Different-Restudy group restudied Block 1-3 items and took a test on Block 4.

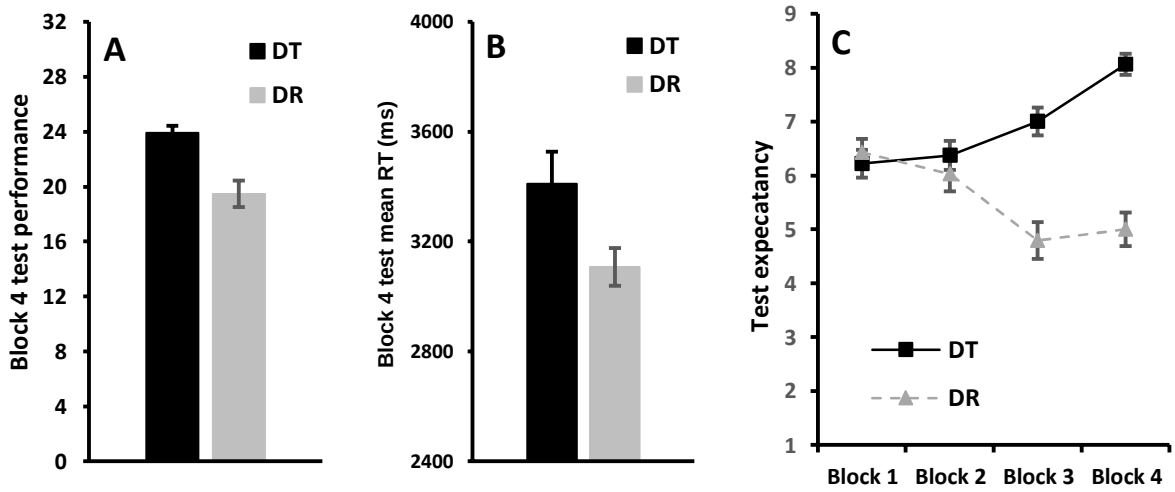

Figure 6. Experiment 3. Panel A: Block 4 test performance; Panel B: Mean RTs in the Block 4 test; Panel C: Test expectancy ratings. DT = Different-Test; DR = Different-Restudy. Error bars represent $\pm$ 1 standard error.

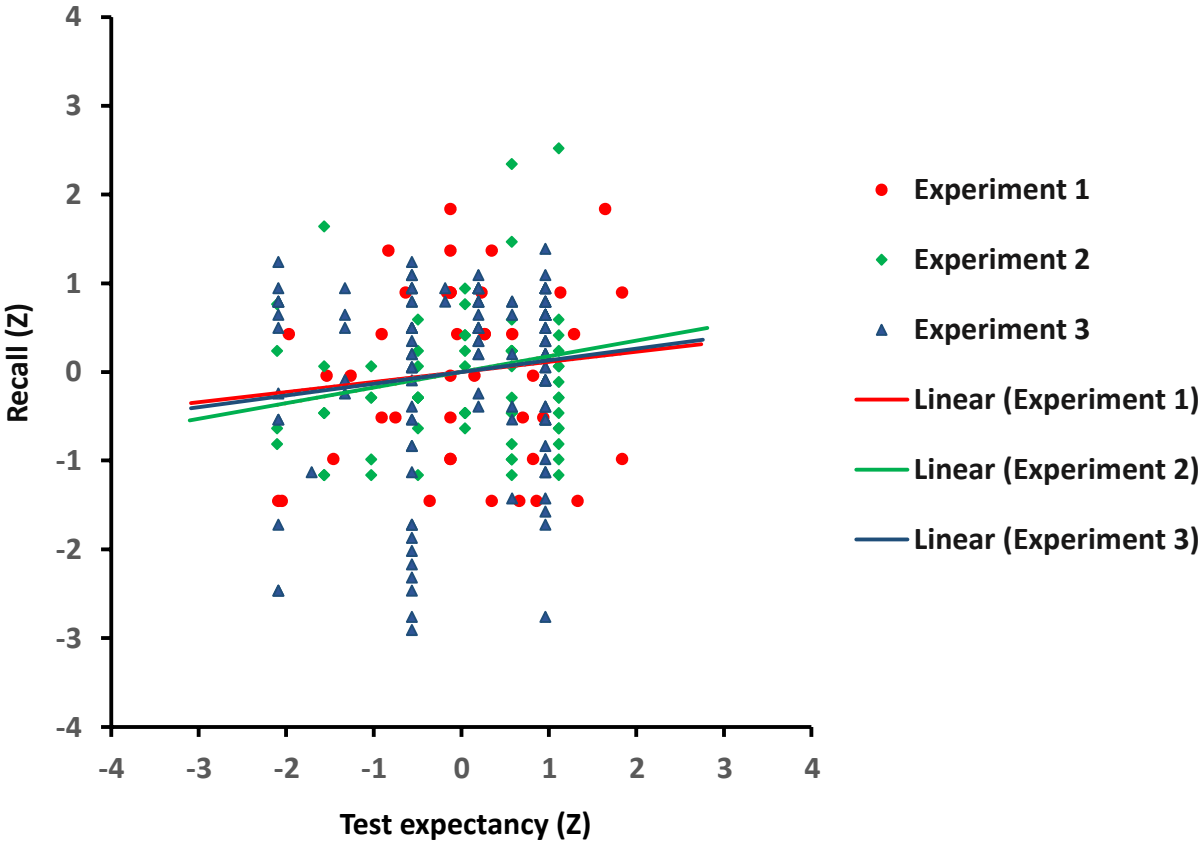

Figure 7. Scatter plot and linear trends between test expectancy and recall in Experiments 1-3 (Experiment 1's Same-Test and Same-Math groups were excluded). Given that the test expectancy rating scales were different, we transformed test expectancy ratings and recall data into *Z* scores in each experiment.

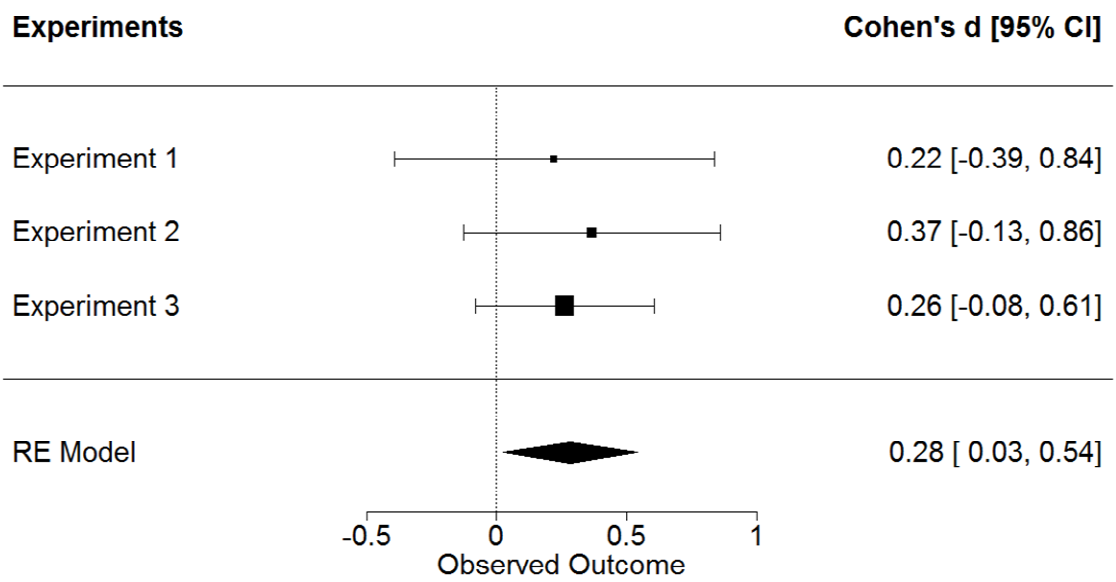

Figure 8. Forest plot of the meta-analysis about the effect of test expectancy on test performance. RE = random effects. Error bars represent 95% CI.

Appendix A

We explored whether classification accuracy (correct *vs* incorrect) moderated the effect of prior interim tests (i.e., Block 1-3 interim tests) on subsequent retrieval effort (i.e., RTs in the Block 4 interim test). For each participant, we calculated mean RTs for correctly classified and incorrectly classified paintings. A mixed ANOVA, with Classification accuracy (correct/incorrect) as a within-subjects variable and Interim task as a between-subjects variable, revealed 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. We warn readers to be cautious about this conclusion 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 groups, revealing the robustness of the difference in retrieval effort between groups and supporting the retrieval-effort theory.