REASARCH REPORT

The Forward Testing Effect: Interim Testing Enhances Inductive Learning

Chunliang Yang and David R. Shanks

University College London

Author Note

This research was supported by the China Scholarship Council (CSC) awarded to Chunliang Yang.

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

All experimental data have been made publicly available via the Open Science Framework (OSF) at https://osf.io/3ga2t/.

Abstract

Induction refers to the process in which people generalize their previous experience when making uncertain inferences about the environment that go beyond direct experience. Here we show that interim tests strongly enhance inductive learning. Participants studied the painting styles of eight famous artists across four lists, each comprising paintings by one pair of artists. In an Interim Test group participants' induction was tested after each list. In two control groups participants solved math problems (Interim Math group) or studied additional new paintings (Interim Study group) following each of Lists 1-3 and were asked to classify new paintings on List 4. In the List 4 interim test, the Interim Test group significantly outperformed the other two groups, indicating that interim testing enhances new inductive learning. In a final cumulative test, accuracy in the Interim Test group at classifying new paintings by studied artists was nearly double that of the other two groups, indicating the major importance of interim testing in inductive learning. This enhancing effect of interim testing on inductive learning was associated with metacognitive awareness.

Keywords: interim testing; inductive learning; metacognitive awareness

Inductive learning is of considerable practical and theoretical interest for learners, educators, and researchers as it is an essential component of how individuals learn and understand the world (Holland, Holyoak, Nisbett, & Thagard, 1989). A substantial body of research has investigated how to improve inductive learning (Djonlagic et al., 2009; Giguere & Love, 2013; Kornell & Bjork, 2008; Mathy & Feldman, 2009; Pashler & Mozer, 2013). However, it is surprising that little research has investigated how to employ testing to enhance inductive learning (Jacoby, Wahlheim, & Coane, 2010), bearing in mind that in the last 100 years, scores of experiments have revealed that repeated testing of studied information enhances its retention more effectively than restudying (Karpicke & Roediger, 2008; Roediger & Karpicke, 2006a).

Testing of previously studied information enhances its learning and retention - the backward testing effect. For example, Roediger and Karpicke (2006b) asked participants to either study a passage four times or study it once and take three tests on it. In a delayed (one-week) test, the repeatedly tested passage was substantially better recalled than the repeatedly studied one. Research has revealed that retrieval practice (i.e., testing) produces deeper and more elaborative learning than restudying and leads to better retrieval accessibility (Roediger & Karpicke, 2006a). Learners can also use test results as feedback to diagnose the gap between their desired and actual learning levels and manage their subsequent learning to narrow the gap (Pyc & Rawson, 2010).

Recent research has shown that testing of studied information also enhances learning and retention of new information (Pastötter & Bäuml, 2014; Yang, Potts, & Shanks, in press-a), which is termed the forward testing effect. For example, Szpunar, Khan, and Schacter (2013) asked participants to study an introductory statistics video which was divided into four sections. Participants either took a test after studying each section, solved math problems following each of sections 1-3 and took a test on section 4, or restudied the preceding section following each of sections 1-3 before taking a test on section 4. Szpunar et al. (2013) found that recall in the section 4 test was substantially better when the preceding three sections had been tested than not tested. Evidence suggests that in the absence of interim testing, more mind wandering occurs during a learning phase (Szpunar, Jing, & Schacter, 2014; Szpunar et al., 2013), and less and less attention (Pastötter, Schicker, Niedernhuber, &

Bäuml, 2011) and learning effort (Yang et al., in press-a) is directed to learning new information. In contrast, when interim tests are administered during a learning phase, subsequent encoding of new information is maintained at the same level or even enhanced compared to the encoding of previous information (Pastötter et al., 2011; Yang et al., in press-a).

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

Research investigating the forward testing effect is largely restricted to low-level learning (Pastötter & Bäuml, 2014). For example, previous research has shown a robust forward testing effect on the learning of face-name pairs (Weinstein, McDermott, & Szpunar, 2011; Yang et al., in press-a), line-drawings of common objects (Pastötter, Weber, & Bäuml, 2013), texts (Wissman, Rawson, & Pyc, 2011), and so on (Jing, Szpunar, & Schacter, 2016; Szpunar et al., 2013; Szpunar, McDermott, & Roediger, 2008). Specifically, participants learned the target items much better if they had been tested rather than untested on previous items. Those studies show that item-level learning is susceptible to enhancement induced by interim testing. It is unknown whether testing can have a facilitatory forward effect on inductive learning. Schacter and Szpunar (2015) suggested that "An important question is

whether interpolated retrieval/testing also enhances learning at a conceptual level" (p. 67). We fill this gap in the present research.

What would we expect to happen if we evaluate the forward testing effect on category induction rather than item learning? There are reasons to predict that category learning will be less enhanced and indeed might even be unaffected by interim testing. Evidence shows that enhancing the encoding of individual exemplars can sometimes have little benefit for category learning. Category induction and stimulus distinctiveness can interact, with induction benefitting much less than identification learning (i.e., item memory) as the stimuli are rendered more distinctive (Love, 2000). For example, Smith, Redford, Washburn, and Taglialatela (2005) studied airport security screeners' ability to detect threatening items in x-ray images. A manipulation which boosted identification of specific items had virtually no effect on generalization to novel exemplars. These findings complement the many other variables known to have divergent effects on exemplar versus category learning, the best-known being the effects of amnesia resulting from temporal lobe damage. Many studies have shown that individuals with amnesia are much more impaired at item memory (recognition) than category induction (Knowlton & Squire, 1993; Reed, Squire, Patalano, Smith, & Jonides, 1999). Theories of category induction have been successful at accounting for these interactions in terms of either the involvement of multiple independent systems underlying the two forms of learning (Ashby, Alfonso-Reese, Turken, & Waldron, 1998) or in terms of the differential demands that induction and identification place on the ability to distinguish stimulus representations (e.g., Nosofsky, Denton, Zaki, Murphy-Knudsen, & Unverzagt, 2012).

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

Experiment 1

Method

Participants

Yang et al. (in press-a) observed effect sizes (i.e., Cohen's *d*s) of the forward testing effect ranging from 0.87 to 1.43. We conducted power analyses using G*power (Faul, Erdfelder, Lang, & Buchner, 2007) and found that about 10-23 participants in each group were required to observe a significant ($\alpha = .05$) forward testing effect at 0.8 power. We tested forty participants, 31 females, with an average age of 21.45 (*SD* = 4.42) years. They were recruited from the University College London (UCL) participant pool and were randomly divided into two equal-sized groups (Interim Test/Interim Math). They gave informed consent and received £4 or course credit as compensation for participating.

Materials

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

Design and procedure

The experiment employs a 2 (Interim task: Interim Test/Interim Math) × 4 (List: 1-4) mixed design, with Interim task as a between-subjects variable and List as a within-subjects variable. Participants were instructed to study the painting styles of various famous painters in anticipation of a cumulative test. They were told that in the first part they would see a list of paintings, consisting of 12

paintings by each of two to-be-studied artists from one set. Following these 24 paintings, they would solve as many math problems as they could in 30 sec (e.g., *47+32 = ?*), and then the computer would decide at random whether or not to give them a short test. If it did, then 12 new paintings (four by each of these two studied artists plus another four by a different artist) would be shown one at a time in a random order and their task was to decide which artist was responsible for each paining. If it did not, they would continue solving math problems for another 60 sec. Then they would go on to the second part identical to the first except that they would learn the styles of two new artists. In fact, the Interim Test group was tested on every list while the Interim Math group was only tested on List 4 (see the design schema in Fig. 1).

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

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

Results and discussion

Fig. 2A shows interim test accuracy. The Interim Test group correctly classified about 65% of paintings (all paintings including those from studied and new artists) and their classification accuracy did not fluctuate across lists, $F(3, 57) = .06$, $p = .98$, $\eta_p^2 < .01$. The Interim Test group correctly classified more List 4 paintings than the Interim Math group, difference = 2.15 paintings, 95% confidence interval (CI) = [.77, 3.53], Cohen's $d = 1.03$, which reveals a substantial forward testing effect.

For the cumulative test, we separately analysed classification of Lists 1-3, List 4, and new artist's paintings, as the Interim Test group underwent an interim test on each of Lists 1-3 but the Interim Math group did not, whereas both groups undertook an interim test on List 4. In the cumulative test, for List 1-3 artists, the Interim Test group correctly classified more paintings than the Interim Math group, difference = 3.35 paintings, 95% CI [1.19, 5.51], d = 0.99 (see Fig. 2B). In addition the Interim Test group correctly classified more List 4 artists' paintings than the Interim Math group, difference = 1.15 paintings, 95% CI [.36, 1.94], d = 0.96, corroborating the pattern found in the List 4 interim test. For new artist, no statistically significant difference was detected between the groups, difference = -0.10 paintings, 95% CI [-.83, .63], d = 0.09. The Interim Math group chose *None of these* somewhat more frequently (M = 8.10, SD = 7.26) than the Interim Test group (M = 6.45, SD = 3.83), although the difference was not statistically significant, difference = 1.65 paintings, 95% CI [-2.07, 5.37], d = 0.34.

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

Experiment 2

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

four from a different artist – the same pictures that were shown in the corresponding test for the Interim Test group) following each of Lists 1-3 and were tested on List 4.

In Experiment 1's cumulative test, there was no difference in classification accuracy for paintings by new artist. There were only four such paintings, and hence it is difficult to explore participants' discrimination between studied and new artists. Therefore, the third change in Experiment 2 was that we added four more paintings by a new artist in the cumulative test. Finally, following study of each list, participants were asked to make a judgment of learning (JOL) by typing in a number (1-9) to indicate their mastery of the two artists' painting styles, and after the completion of List 4, participants were also asked to rate their mastery of all 8 artists' painting styles.

Method

Participants

Seventy-two participants, 45 females, with an average age of 25.44 (*SD* = 7.32) years, were recruited from the UCL participant pool and were randomly divided into three groups (Interim Test/Interim Math/Interim Study). They gave informed consent and received £4 or course credit as compensation for participating.

Materials, design, and procedure

The same paintings plus another four by Jan Brueghel the Elder were employed. Experiment 2 involved a 3 (Interim task: Interim Test/Interim Math/Interim Study) × 4 (List: 1-4) mixed design. The procedure was the same as in Experiment 1 with the following exceptions. Participants were informed that, after studying 24 paintings and solving math problems for 30 sec, the computer would randomly decide the following task. If it decided to give them a short test, 12 new paintings would be presented one at a time and their task was to classify each painting. If it decided to give them more math problems, they would continue solving math problems for another 60 sec. If it decided to give them more new paintings, 12 new paintings (four by each of two studied artists and four by a different artist) would be presented with the artist's name or *None of these* displayed below, one at a time for 5 sec in a random order. In fact, the Interim Test group was tested on every list. The Interim Math group

continued solving math problems following each of Lists 1-3 and was tested on List 4. The Interim Study group studied 12 new paintings following each of Lists 1-3 and was tested on List 4 (see the design schema in Fig. 1).

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

Results and discussion

Fig. 2C shows interim test classification. The Interim Test group's classification accuracy did not fluctuate across lists, $F(3, 69) = .22$, $p = .88$, $\eta_p^2 = .01$. For the List 4 interim test classification, a one-way ANOVA showed a main effect of Interim task, $F(2, 69) = 4.85$, $p = .01$, $\eta_p^2 = .12$. Again revealing a forward testing effect, the Interim Test group correctly classified more List 4 paintings than the Interim Math group, difference = 2.13 paintings, 95% CI [.51, 3.74], $d = 0.78$, and more than the Interim Study group, difference = 2.17 paintings, 95% CI [.58, 3.75], $d = 0.81$, but there was no significant difference between the Interim Study and Interim Math groups, difference = -0.04 paintings, 95% CI [-1.64, 1.56], $d = 0.02$.

Fig. 2D shows cumulative test classification. For List 1-3 artists, a one-way ANOVA revealed a main effect of Interim task, $F(2, 69) = 9.35$, $p < .001$, $\eta_p^2 = .21$. The Interim Test group correctly classified more paintings than the Interim Math group, difference = 3.42 paintings, 95% CI [1.53, 5.30], $d = 1.08$, and more than the Interim Study group, difference = 3.08 paintings, 95% CI [1.19,

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

4.98], d = 0.97, but there was no significant difference between the Interim Study and Interim Math groups, difference = 0.33 paintings, 95% CI [-1.11, 1.78], d = 0.13. For the List 4 artists, a one-way ANOVA showed a main effect of Interim task, $F(2, 69) = 3.81$, p =.03, η_p^2 = .10. The Interim Test group correctly classified more paintings than the Interim Math group, difference = 1.04 [.23, 1.86] paintings, d = .76, and more than the Interim Study group, difference = 1.08 paintings, 95% CI [.20, 1.96], d = 0.73, but there was no significant difference between the Interim Study and Interim Math groups, difference = -0.04 paintings, 95% CI [-.79, .71], d = 0.03. These results corroborate the pattern in the List 4 interim test.

For new artists, a one-way ANOVA showed a main effect of Interim task, $F(2, 69) = 3.21$, $p < .05$, η_p^2 = .09. The Interim Test group correctly classified more new artists' paintings than the Interim Math group, difference = 1.13 paintings, 95% CI [.01, 2.24], d = 0.60, and more than the Interim Study group, difference = 1.33 paintings, 95% CI [.14, 2.53], d = 0.66, but there was no significant difference between the Interim Study and Interim Math groups, difference = -0.21 paintings, 95% CI [-1.31, .90], d = 0.11. These results indicate that the Interim Test group was better able to discriminate studied from new artists' paintings than the other two groups.

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

For List 4 JOLs, a one-way ANOVA revealed a main effect of Interim task, $F(2, 69) = 4.46$, p = .02, η_p^2 = .11. List 4 JOLs in the Interim Test group were higher than those in the Interim Math

group, difference = 1.08, 95% CI [.16, 2.01], d = 0.69, and higher than those in the Interim Study group, difference = 1.50, 95% CI [.46, 2.54], d = 0.86, but there was no significant difference between the Interim Study and Interim Math groups, difference = -0.42, 95% CI [-1.57, .74], d = 0.21. These results reveal that List 4 JOLs were aligned with List 4 interim test classification.

For global JOLs, a one-way ANOVA showed a main effect of Interim task, $F(2, 69) = 4.40$, $p = .02$, $\eta_p^2 = .11$. Global JOLs in the Interim Test group were higher than those in the Interim Math group, difference = 1.21, 95% CI [.34, 2.08], d = 0.82, and higher than those in the Interim Study group, difference = 1.21, 95% CI [.27, 2.14], d = 0.77, but there was no significant difference between the Interim Study and Interim Math groups, difference = .00, 95% CI [-1.03, 1.03]. These results reveal that global JOLs aligned with cumulative test classification.

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

In Experiment 2, List 4 interim test classification reveals that interim testing enhances inductive learning of new categories more effectively than no interim testing or studying more new exemplars – a forward testing effect. In the cumulative test, the Interim Test group was better able to classify studied artists' paintings and discriminate between the paintings of studied and new artists than the other two groups. The Interim Math and Interim Study groups recognized the reduction in their inductive learning across lists whereas the Interim Test group was aware of the maintenance of their learning across lists.

In the List 4 interim test as well as in the cumulative test, the Interim Study group failed to classify paintings any better than the Interim Math group. This might seem surprising given that the Interim Study group had the opportunity at the end of each list to study four additional paintings by the two target artists. However this lack of benefit of additional study opportunities is in line with many comparable findings in the backward testing effect (e.g., Roediger, Agarwal, McDaniel, &

McDermott, 2011) and the rereading effect (e.g., Callender & McDaniel, 2009) literatures. This finding serves to emphasize that the benefit of testing seen in the Interim Test group is not simply due to additional exposure to relevant learning materials. It is the act of being tested on one's knowledge that causes the benefit.

General discussion

Previous research has shown that many variables have stronger effects on item (exemplar) versus inductive (category) learning (Knowlton & Squire, 1993; Love, 2000). However, contrary to the hypothesis that interim testing might have a smaller effect on inductive learning than it has on item memory (or even no facilitatory effect), both of the experiments reported here clearly reveal a robust forward testing effect, a finding which has not previously been demonstrated. A few factors may contribute to the facilitatory effect of interim testing on inductive learning.

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

Pastötter et al. (2011) proposed a reset of encoding theory to explain the forward testing effect which may operate as well as or instead of the aforementioned motivational mechanisms. Pastötter et

al. suggested that interim testing causes internal context change between successive lists, which resets subsequent encoding of new information and renders it as effective as encoding of prior information. Evidence for this mechanism comes from a study by Pastötter, Bäuml, and Hanslmayr (2008). These researchers measured participants' brain oscillatory activity while encoding two lists of words. Participants were instructed to either perform an imagination task or not following encoding of the first list, and then studied the second list. Alpha (8-14 Hz) and theta (4-7 Hz) power (synchrony in brain oscillations), which are linked to reduced attention, increased across lists if participants did not perform the imagination task. The inference is that the imagination task produced an internal context change between the lists, and this context change attenuated the increase in alpha and theta power that would otherwise have occurred. Thus internal context change between lists resets the encoding of new information and makes it as effective as prior encoding.

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

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

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

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

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

In Experiment 2, global JOLs were made following the List 4 interim test and hence might be affected by List 4 interim test classification. The Interim Test group outperformed the other two groups in the List 4 interim test, which may have induced them to report higher global JOLs than the other two groups. To test this idea, we explored the correlation between List 4 interim test classification and global JOLs at the participant level. Consistent with this idea, the correlation was positive, $r = .26$, $F(1, 71) = 5.21$, $p = .03$, $R^2 = .07$, adjusted $R^2 = .06$. Anchoring may provide another

possible mechanism: the Interim Test group gave higher list-by-list JOLs to List 4 than the other two groups and these JOLs might act as anchors for global JOLs, driving higher global JOLs in the Interim Test group than in another two groups. To test this idea, we explored the correlation between List 4 JOLs and global JOLs at the participant level. There was a positive correlation, $r = .81$, $F(1, 71) = 134.67$, $p < .001$, $R^2 = .66$, adjusted $R^2 = .65$.

In conclusion, interim testing enhances inductive learning more effectively than no interim testing or studying more new exemplars – the forward testing effect on inductive learning. This forward testing effect is associated with metacognitive awareness. The present research found a forward testing effect and Jacoby et al. (2010) found a backward testing effect on inductive learning. Collectively, these findings lead to a strong recommendation that interim testing should be employed to enhance inductive learning in the classroom and elsewhere.

References

Anderson, J. R. (2000). *Learning and Memory*. New York: Wiley.

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

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

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

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

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

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

Holland, J. H., Holyoak, K. J., Nisbett, R. E., & Thagard, P. R. (1989). *Induction: Processes of Inference, Learning, and Discovery.* Cambridge: MIT Press.

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

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

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

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

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

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

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

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

Murphy, G. L. (2002). *The Big Book of Concepts*. Cambridge: MIT Press.

Nosofsky, R. M., Denton, S. E., Zaki, S. R., Murphy-Knudsen, A. F., & Unverzagt, F. W. (2012). Studies of implicit prototype extraction in patients with mild cognitive impairment and early Alzheimer's disease. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 38*(4), 860-880. doi: 10.1037/a0028064

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

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

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

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

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

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

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

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

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

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

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

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

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

Smith, J. D., Redford, J. S., Washburn, D. A., & Taglialatela, L. A. (2005). Specific-token effects in screening tasks: Possible implications for aviation security. *Journal of Experimental*

Psychology: Learning, Memory, and Cognition, 31(6), 1171-1185. doi: 10.1037/0278-7393.31.6.1171

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

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

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

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

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

Weinstein, Y., McDermott, K. B., & Szpunar, K. K. (2011). Testing protects against proactive interference in face-name learning. *Psychonomic Bulletin & Review, 18*(3), 518-523. doi: 10.3758/s13423-011-0085-x

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

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

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

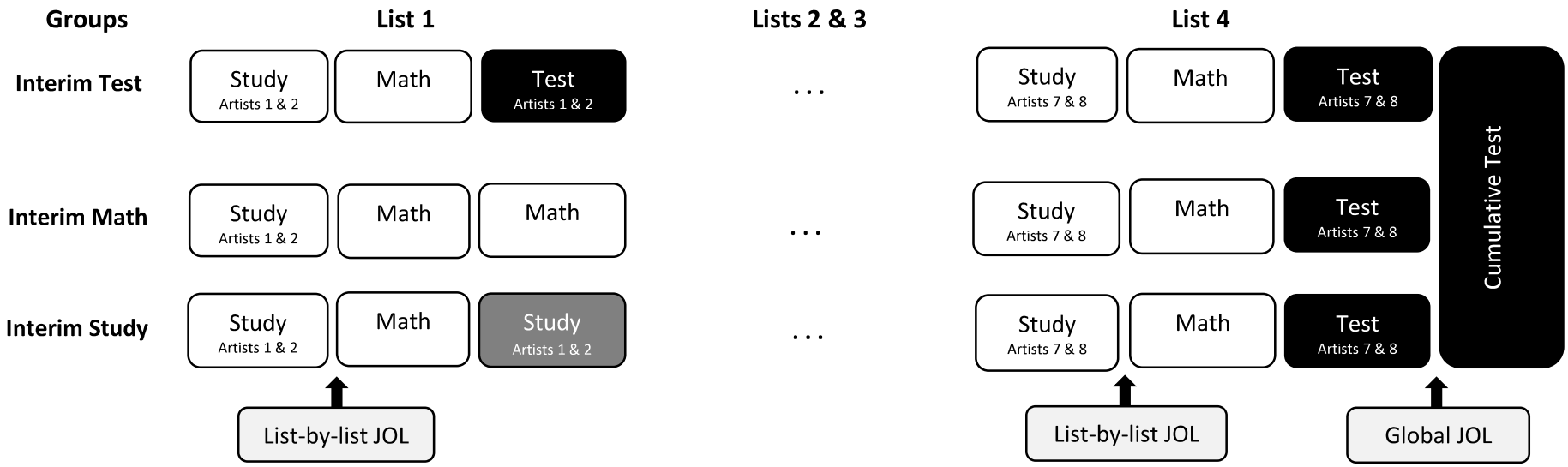

Fig. 1. Experimental design schema for the Interim Test, Interim Math, and Interim Study groups. Lists 2 and 3 were identical to List 1 in each group, but with different artists. The Interim Study group was not included in Experiment 1. Judgments of learning (JOLs; i.e., metacognitive judgments about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

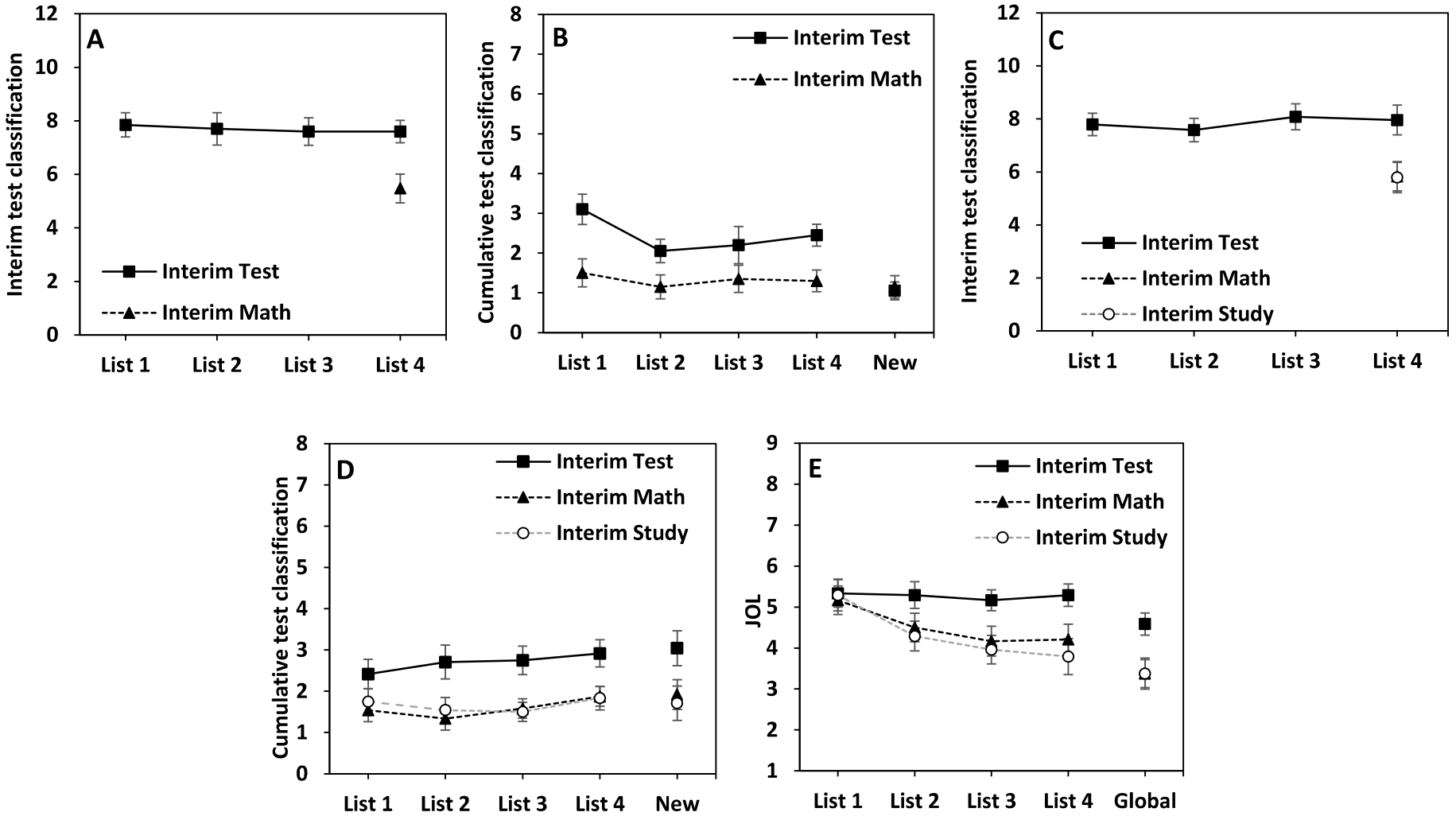

Fig. 2. A: Interim test classification accuracy (no. correct) across lists in Experiment 1. B: Cumulative test classification accuracy (no. correct) across lists and for one new artist (*None of these*) in Experiment 1. C: Interim test classification accuracy (no. correct) across lists in Experiment 2. D: Cumulative test classification accuracy (no. correct) across lists and for two new artists (*None of these*) in Experiment 2. E: List-by-list and global JOLs in Experiment 2. Error bars represent ± 1 standard error.

Note: In Panels B and D, classification accuracy in the cumulative test is broken down according to the artists who appeared in each list. For example, the data at "List 1" are the numbers of correct classifications of the eight new paintings by Artists 1 and 2. Chance performance for each list given nine response categories (eight studied artists and *None of these*) is therefore 8 × 1/9 = 0.89 correct responses. In Panel B, chance performance for one new artist (*None of these*) is 4 × 1/9 = 0.44, and, in Panel D, chance performance for two new artists is 8 × 1/9 = 0.89 correct responses.