Educational Psychology Review (2023) 35:51
https://doi.org/10.1007/s10648-023-09772-y

INTERVENTION STUDY

When More Is Not Better: Effects of Interim Testing and Feature Highlighting in Natural Category Learning

Yewon Kang[1,2] · **Hyorim Ha**[1] · **Hee Seung Lee**[1]

Accepted: 15 April 2023 / Published online: 27 April 2023

Abstract

Natural category learning is important in science education. One strategy that has been empirically supported for enhancing category learning is testing, which facilitates not only the learning of previously studied information (backward testing effect) but also the learning of newly studied information (forward testing effect). However, in category learning, such benefits of testing have mostly been examined without explicit instructions. This is not aligned with a real educational practice where teachers often provide students with explicit instructions that highlight the diagnostic features of the category. Thus, we investigated the effect of interim testing and feature highlighting on rock category learning and whether the provision of feature highlighting further enhances testing benefits. The participants learned 12 rock categories, which were divided into two sections (Sections A and B). They studied a series of rock images with or without feature highlighting and were given an interim test or not on Section A before proceeding to Section B. After Section B, all the participants took a final test in which they had to classify both old and new rock images of the studied categories. Three experiments demonstrated the benefits of interim testing (compared to restudy) for both previously and newly studied categories. However, feature highlighting did not further enhance learning and sometimes even impeded learning. The findings suggest that providing more information is not always better than providing less information in natural category learning.

Keywords Natural category learning · Backward effect of testing · Forward effect of testing · Interim testing · Feature highlighting

✉ Hee Seung Lee
hslee00@yonsei.ac.kr

[1] Department of Education, Yonsei University, 50 Yonsei-Ro, Seodaemun-Gu, Seoul, Korea

[2] Present Address: Department of Educational Psychology, University of Minnesota, Minneapolis, MN, USA

Natural categories are widely taught in science classes; for example, students learn about rock types in a geology class and tree species in a biology class. When studying categories, students should draw similarities and differences between exemplars and classify them based on their features. In a biology class, for example, trees with oval leaves should be grouped together, whereas those with linear leaves should be classified as different species. For successful category learning, students should grasp the characteristics of the exemplars within the same category and make inferences about new exemplars (Murphy, 2002). However, despite its importance in science education, only a few attempts have been made to optimize natural category learning (for a review, see Nosofsky & McDaniel, 2019).

One strategy that has been empirically supported for enhancing category learning is testing. A large number of studies demonstrated the robust effect of testing relative to a restudy control, using a wide range of materials, including verbal and nonverbal (for reviews, see McDermott, 2021; Rowland, 2014). The effect has been consistently shown not only in lab settings but also in classrooms (Schwieren et al., 2017; Sotola & Crede, 2021; Yang et al., 2021). In the context of category learning, early research revealed testing benefits by comparing classification performance between observational training and feedback training (Ashby et al., 2002; Carvalho & Goldstone, 2015). Observational training is similar to a restudy condition in that an exemplar is presented with its category label whereas feedback training is similar to a test condition in that an exemplar is presented without the category label, requiring participants to classify it. Such research has shown that feedback training is superior to observational training (e.g., Edmunds et al., 2015), suggesting that the benefits of testing may extend to category learning. To our knowledge, the first study that directly examined testing effects in category learning was Jacoby et al.'s study (2010). They compared the effects of repeated testing and repeated study on natural category learning using bird families as study materials. When participants were asked to categorize the studied and novel exemplars into the corresponding bird families, the repeated testing group outperformed the repeated study group in both recognition memory and transfer classification tasks. This study showed that testing on the studied categories can enhance learning of those tested categories. For successful classification, students should discover features that specify which category the exemplars fall under and retrieve critical information from previously studied exemplars that can later be used to classify novel exemplars (Nosofsky & McDaniel, 2019). The testing group had to repeatedly retrieve the category label and feature, which likely increased attention to the studied exemplars (Jacoby et al., 2010). This can facilitate the memory of the category label and features, and possibly the link between them, resulting in better classification performance (cf. Lee & Ahn, 2018).

The positive effect of testing can extend to subsequent learning of new categories. Recently, emerging research has shown that testing facilitates not only the learning of the tested information but also the learning of the novel information studied after the test. Pastötter and Bäuml (2014) referred to the former as *the backward effect of testing* and the latter as *the forward effect of testing*. Particularly in category learning, several studies have demonstrated the forward effect of testing using painting styles (Choi & Lee, 2020; Lee & Ahn, 2018; Lee & Ha, 2019; Yang & Shanks, 2018; Yang et al., 2019). For example, Lee and Ahn (2018) investigated how testing

on previously learned paintings of multiple artists affected the learning of new paintings of different artists. In their study, participants learned six different painting styles in the first section and the other six styles in the second section. Before the second section, participants were given either an interim test or an interim restudy on the paintings from the first section. In the final classification test, although all participants learned under the same conditions in the second section, the participants who took an interim test showed better classification for the paintings from the second section than those who had an interim restudy, indicating the forward effect of testing.

There are several accounts that explain the forward effect of testing (for reviews, see Chan et al., 2018b; Yang et al., 2018). Some researchers suggested that testing facilitates learning by increasing test expectancy (i.e., "I will be tested soon."). While taking an interim test on initial learning, students can increase their expectancy about an upcoming test and put greater effort into their subsequent learning, resulting in improved performance on a later test (Weinstein et al., 2014; Yang et al., 2017). Such test expectancy has been shown to encourage students to spend more time studying (Yang et al., 2017) and reduce mind wandering (Jing et al., 2016; Szpunar et al., 2013). Some other researchers suggested that testing can reduce proactive interference among different study sessions (Szpunar et al., 2008). Interpolated tests can create contextual segregation between encoding and retrieval by isolating the retrieved materials from the subsequently studied materials (Davis et al., 2017; Kliegl & Bäuml, 2021). Also, the forward testing effect can occur via metacognitive benefits. While being tested, students can evaluate and adjust their learning strategies (Cho et al., 2017). Indeed, Chan et al. (2018a) demonstrated that previously tested participants used a more effective retrieval strategy (semantic organization) than no-test and restudy conditions. These non-mutually exclusive accounts suggest that testing can enhance natural category learning in several ways. For example, during an interim test, students may realize that classification is more difficult than they expected, thus putting more effort into their subsequent learning. Also, they may realize that noticing similarities among exemplars within categories is not sufficient for successful classification and that they need to identify differences that distinguish one category from another (Birnbaum et al., 2013; Kang & Pashler, 2012), resulting in strategy change in subsequent learning.

Although there is strong evidence that testing is an effective way to improve category learning, one of the major limitations of prior research is that the testing effects were only examined in a situation where learners had to identify features that are relevant to categorization on their own by studying a series of examples. That is, participants were not provided with any feature descriptions (e.g., Jacoby et al., 2010; Lee & Ahn, 2018; Yang & Shanks, 2018); instead, they had to abstract rules and generalize what they had learned to other new instances for themselves. However, such abstraction and generalization may not be successful. Particularly in natural category learning, exemplars often include ambiguous properties with no clear separation among categories (McCloskey & Glucksberg, 1978; Murphy, 2002). Thus, it may be challenging for students to identify key features on their own, exacerbating the difficulty when they have little prior knowledge. For successful category learning, students should develop and test rules that determine category membership

(Markant, 2019), and such hypothesis search and testing can increase the working memory burden. Especially, when taking a test (rather than restudying the material), students should retrieve the relevant information, which can add more to the cognitive burden. Besides, they have to process feedback that requires additional attention and effort. Such excessive cognitive load may interfere with learning (Sweller, 2010; Sweller et al., 2019) and undermine the benefits of testing (Leahy & Sweller, 2019). Consistent with this argument, van Gog and Sweller (2015) also claimed that the testing effect decreases as the complexity of learning material increases (but see also Karpicke & Aue, 2015). This suggests that one possible way to increase the benefits of testing is to explicitly provide students with key features that are relevant to successful categorization. For example, when learning rock categories, students can be given feature descriptions of the rock images. This strategy is called *feature highlighting* (Eglington & Kang, 2017; Miyatsu et al., 2019). We predict that, when learners are given feature highlighting, because they do not have to discover classification rules on their own, they may be able to allocate their cognitive resources to activities that are more relevant to learning rather than inefficient search and hypothesis testing, thereby increasing the benefits of testing.

The direct evidence of the effect of feature highlighting comes from a study by Miyatsu et al. (2019). In their study, participants classified rock images and received feedback with or without feature descriptions in classification trials. The control group received only the category name for each rock image as feedback, whereas the feature-description group received the same image but with the feature descriptions (e.g., "rounded fragments cemented together" circled in blue) on the corresponding part of the image. The results showed that the feature-description group classified novel rock exemplars better than the control group.

Feature highlighting can help learners recognize critical rules, which is a necessary step in category learning. According to the competition between verbal and implicit systems (COVIS) model (Ashby et al., 1998), students initially develop explicit rules that are easy to verbalize during hypothesis testing. Similarly, the rule-plus-exception model (RULEX) proposes that simple rules are constructed first, with certain exceptions maintained in memory (Nosofsky et al., 1994). Furthermore, learners tend to employ rule-based strategies when there is no explicit learning such as feedback (Ashby et al., 1999). Feature highlighting, which provides explicit verbal descriptions of key category features, can be an effective method for identifying simple rules. When explicit rules are provided, students may learn more efficiently by reducing the time required to construct hypotheses and decide whether to accept them (Miyatsu et al., 2019), resulting in better use of their working memory and attention (Zeithamova & Maddox, 2006).

Including feature highlighting in learning materials might also be a more applicable strategy in real-world contexts. As previously stated, most prior studies on category learning required learners to extract rules and features that are useful for classification on their own, without providing explicit feature descriptions. Nevertheless, in educational contexts, learning is intended from the outset—teachers often inform students of the category features from the beginning phase of a lesson. For example, in a geology class, it is common in textbooks to include verbal descriptions of the to-be-learned rock categories, such as "sandstone is composed of sand-sized grains"

and "breccia is composed of angular fragments." Such discrepancies between the research procedure and actual educational practice make it difficult to predict the benefit of feature highlighting and its generalizability to real educational settings.

Moreover, in class, teachers often combine several tactics rather than using a single effective strategy. Thus, we aimed to examine whether interim testing and feature highlighting would facilitate learning and whether these two strategies would lead to an additive effect when combined. A combination of effective strategies may not always boost learning outcomes. While some previous studies suggested that the effectiveness of testing can be increased with additional strategies (e.g., Cummings et al., 2022; Rawson & Dunlosky, 2011), other studies provided counterevidence (e.g., Kubik et al., 2020; Miyatsu & McDaniel, 2019; O'Day & Karpicke, 2021; van den Broek et al., 2021). Therefore, although feature highlighting is a widely used technique in the classroom, it is unclear how it interacts with another effective learning strategy—testing. To the best of our knowledge, no prior research examined the combined effects of interim testing and feature highlighting. This study aimed to investigate whether the testing effects (both the backward and forward effects of testing) occur in natural category learning and whether feature highlighting further increases the benefits of testing.

The Present Study

The present research chose rock categories as learning materials to examine the effects of interim testing and feature highlighting on natural category learning. Several studies have successfully used rock categories to investigate the effect of feature highlighting in natural category learning (e.g., Meagher et al., 2022; Miyatsu et al., 2019). Figure 1 shows a schematic representation of the procedures of three experiments. To examine both backward and forward effects of testing, we employed the forward-testing effect test procedure (Lee & Ahn, 2018), in which participants were asked to either take a test or not between two learning sections. In the present study, participants studied 12 rock categories that were divided into two learning sections (Sections A and B). After studying Section A, they either took an interim test or restudied the rock categories learned in Section A and then completed Section B. Therefore, the experimental manipulation of interim activity occurred only for Section A (i.e., interim test or interim restudy), and all participants studied Section B under the same conditions. After Section B, all participants took a final test on both learning sections. To avoid the interim test conditions benefiting from the identical test format, the interim test was a cued-recall test whereas the final test was a multiple-choice classification test. Moreover, for successful category learning, students should be able to transfer what they have learned to new exemplars. Thus, in all three experiments, we included both studied exemplars (retention) and novel exemplars (transfer) in the final test. If the participants who took an interim test scored higher on Section A than those who did not, it was viewed as the backward effect of testing. If the participants who took an interim test scored higher on Section B than those who did not, it was viewed as the forward effect of testing.

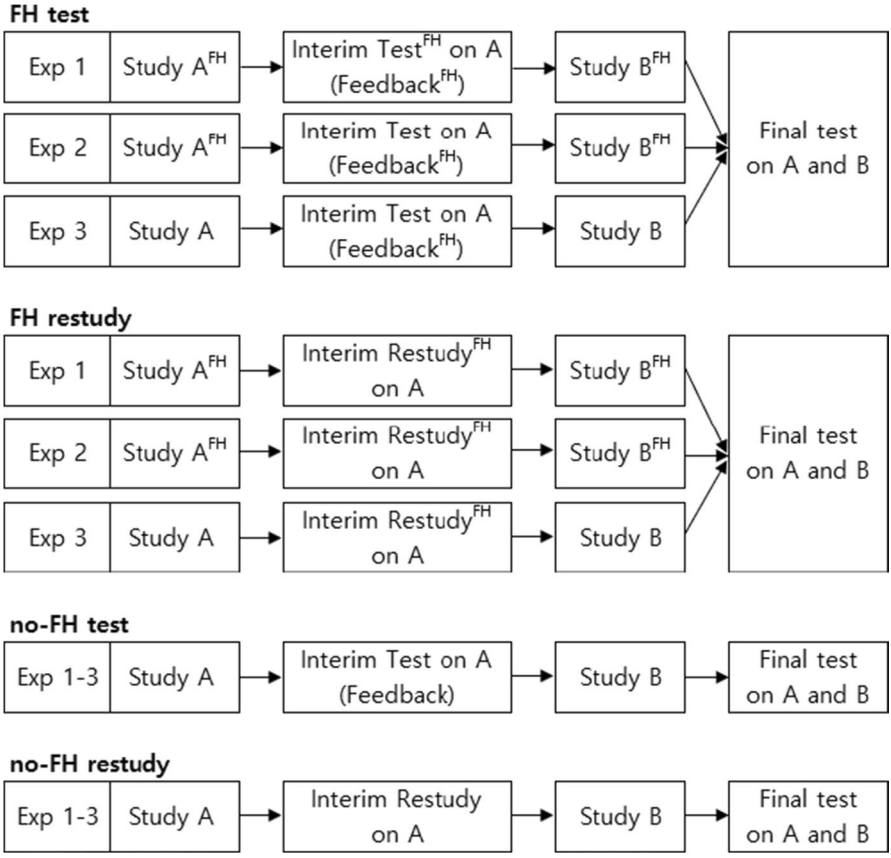

Fig. 1 Schematic representation of the procedures used in Experiments 1–3. The superscript FH represents feature highlighting. The no-FH test and no-FH restudy conditions were identical across the three experiments

Additionally, we investigated whether feature descriptions promote natural category learning by manipulating the presence of feature highlighting. As the experiment progressed, we changed when and how feature highlighting was presented. In Fig. 1, the superscript FH indicates that feature highlighting was provided. Across the three experiments, no-FH conditions were identical; we varied only FH conditions. In Experiment 1, feature highlighting was provided during the entire learning (both Sections A and B) and interim test sessions (both on the test problem images and feedback). In Experiment 2, feature highlighting was presented throughout the learning session (both Sections A and B) and as feedback (but not on the test problem images) during the interim test session. In Experiment 3, feature highlighting was only presented as feedback during the interim test session.

Experiment 1

Method

Participants

Participants were recruited from a large university in South Korea via school-wide online advertisements. To determine the sample size, we conducted an a priori power analysis using G*Power (Faul et al., 2007). Miyatsu et al., (2019, Experiment 2) reported a very large benefit of feature highlighting. To detect a medium (η_p^2 =0.06) to large-sized (η_p^2 =0.14) effect of the condition, a required sample size was between 17 and 42 per condition for a 2 (interim activity: test vs. restudy)×2 (feature highlighting: FH vs. no-FH) between-subjects factorial design at a power of 0.9. Thus, in Experiments 1 and 2, recruitment continued until reaching 25 participants per condition after excluding some participants based on the exclusion criteria. In Experiment 1, 119 undergraduates participated online in return for monetary compensation. We excluded two participants whose test scores were above or below the three *SD* of the group mean, 13 participants who reported a high level of prior knowledge in rock categories, and four participants who left the experiment idle on the self-paced page for longer than 10 min. Thus, there were 100 participants (68 women; 32 men; mean age = 22.40 years) for the final analyses. Each participant was randomly assigned to one of four conditions: 25 in the feature highlighting test (FH test), 25 in the no feature highlighting test (no-FH test), 25 in the feature highlighting restudy (FH restudy), and 25 in the no feature highlighting restudy (no-FH restudy). All experiments were approved by the Institutional Review Board (IRB) of the university where the experiments were conducted.

Design

A 2 (interim activity: test vs. restudy)×2 (feature highlighting: FH vs. no-FH) between-subjects factorial design was employed. Depending on the experimental condition, participants were asked to take a test or not (i.e., restudy) during the interim activity session between Sections A and B. Additionally, during the learning and interim activity sessions, category features were either presented or not.

Materials

The rock images used in this study were adapted from Miyatsu et al. (2019) (retrieved from https://osf.io/9vg8m/), which were originally created by Nosofsky et al. (2018). There were 12 rock categories: amphibolite, breccia, conglomerate, gneiss, granite, marble, obsidian, pegmatite, pumice, rock gypsum, sandstone, and slate. In the FH conditions, the characteristic feature descriptions were presented on the corresponding parts of each rock image with colored circles, and the category name was shown below the rock image. For example, for marble, the feature

descriptions "grey to white crystalline material" and "darker swirls and veins" were included in the corresponding parts of the image along with the category name below the rock image. This feature highlighting was done following the spatial contiguity principle, which refers to placing text near corresponding parts of an image or animation (Mayer, 2005) to minimize gaze shifts between images and feature descriptions (Miyatsu et al., 2019). In the no-FH conditions, only the category name was shown below the rock image. All materials were presented in Korean, including the names of rock categories and feature descriptions.

Procedure

Before Experiment 1, we conducted a pilot experiment with undergraduate students ($N=5$) to determine the appropriate number of study trials and study time. All the materials and procedures were prepared according to this pilot and relevant prior research. Participants took part in an online experiment via Qualtrics (https://www.qualtrics.com/) from their preferred location. After receiving initial instructions on the purpose and overall procedure of the experiment, all participants were told that they would learn 12 different rock categories divided into two sections (Sections A and B). To control for general test expectancy at the beginning of the experiment, all participants were informed that they would be given a test where they would have to classify old and new rock images from the studied rock categories.

All participants learned six of the 12 rock categories in Section A and the other six in Section B. Each category has six exemplars, for a total of 72 rock images. The rock-section pairs were counterbalanced between the two sections to prevent the specific-item effect. In Section A, participants studied 36 rock images, six from each of the first six rock categories. The rock images were interleaved (Kornell & Bjork, 2008) and presented one by one in a fixed random order. Each image was displayed for 4 s. In the FH conditions, the name of the rock and feature highlighting were embedded in each rock image. In the no-FH conditions, only the name of the category was presented below each rock image. Following Section A, participants were given a different interim activity. They either took an interim test on the same 36 rock images studied in Section A or restudied them. In the interim test conditions, participants took a cued-recall test of the Section A material. They were given the same rock images studied in Section A without the category names; they had to enter the name of the corresponding rock category for each image at their own pace. Each trial was followed by immediate feedback for 2 s. The feedback page was identical to that used in the learning session of Section A. In the interim restudy conditions, participants restudied the same rock images that they had previously studied in Section A. Each image was presented for 6 s in the same manner as in the learning session of Section A. After finishing the interim activities for Section A, participants proceeded to Section B, where they studied 36 rock images from another six rock categories, each with six images. The rock images were presented in the same manner as in Section A. Upon completion of Section B, all participants were given a final test on both sections. The final test was a multiple-choice test to evaluate participants' ability to correctly classify exemplars into categories. It included two old and

two new rock images from each of the 12 studied categories, resulting in a total of 48 test items. For each test trial, a rock image was presented, and participants had to select the corresponding category name among the 12 different rock categories. The rock images were not separated by sections; they were interleaved and presented in a fixed random order. Thus, participants did not know which section of rock categories each image belongs to. The final test was self-paced, and feedback was not provided. After completing the final test, participants were debriefed and thanked.

Results

For all primary analyses, we reported data for both null hypothesis significance testing and inclusion/exclusion Bayes factors, which encode the Bayes Factor for each predictor across all matched models. Because the present research has multiple predictors (i.e., interim activity, feature highlighting, and test item type), the number of alternative models grows too large when adding factors. Thus, we computed the model-averaged results and reported the inclusion Bayes Factor (i.e., BF_{incl}) when p-values are significant, and the exclusion Bayes Factor (i.e., BF_{excl}) when p-values are not significant. The analyses were performed using JASP (JASP Team, 2022), and Jeffreys' scheme was used to interpret the strength of evidence. A Bayes factor of 1 to 3 was considered as anecdotal evidence, 3 to 10 as moderate evidence, 10 to 30 as strong evidence, 30 to 100 as very strong evidence, and more than 100 as decisive evidence for a model including or excluding a particular predictor term (Jeffreys, 1961; see also Lee & Wagenmakers, 2014).

Interim Test Performance

Only the participants in the FH and no-FH test conditions were given an interim test. The mean interim test score was 91.11 ($SD = 8.29$) in the FH test condition and 87.67 ($SD = 8.86$) in the no-FH test condition. The mean difference between the FH and no-FH test conditions was not significant, $t(48) = 1.42$, $p = 0.162$.

Final Test Performance

Figure 2 shows the mean percentage of correct classification for the old (retention) and new rock images (transfer) in the final test of Sections A and B. To investigate the forward and backward effects of testing, we conducted two separate $2 \times 2 \times 2$ mixed analyses of variance (ANOVAs) on the percentage of correctly classified rock images for Sections A and B. Feature highlighting (FH vs. no-FH) and interim activity (test vs. restudy) were included as between-subjects factors, while the test item type (old vs. new) was included as a within-subject factor.

To investigate the backward effect of testing, we first examined the performance for Section A. There was a significant main effect of test item type, $F(1, 96) = 136.43$, $p < 0.001$, $\eta_p^2 = 0.587$, $BF_{incl} = 3.95 \times 10^{+16}$, indicating decisive evidence for the inclusion of the test item type in the model. The percentage of correctly classified

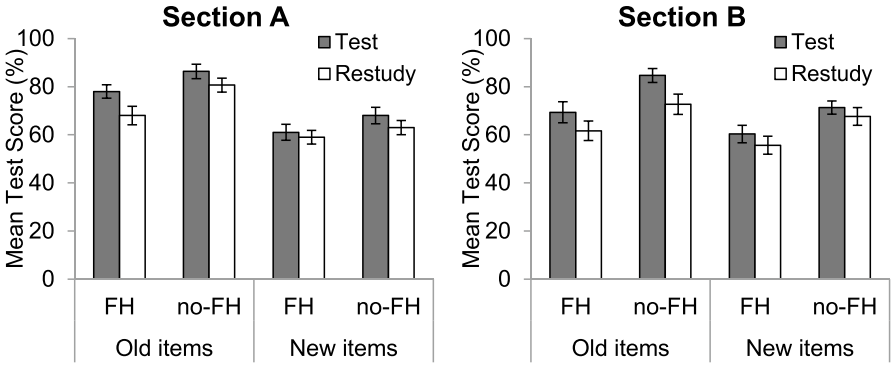

Fig. 2 Mean percentage of the correct responses for the old and new items in the final test of Sections A and B in the FH test, no-FH test, FH restudy, and no-FH restudy conditions of Experiment 1. Error bars represent one standard error of the mean

rock images was significantly higher for the old items ($M = 78.25$, $SD = 16.83$) than for the new items ($M = 62.75$, $SD = 15.82$), $d = 1.14$. There was a significant main effect of interim activity, $F(1, 96) = 3.97$, $p = 0.049$, $\eta_p^2 = 0.040$, $BF_{incl} = 1.64$, indicating anecdotal evidence for the inclusion of the interim activity in the model. The participants who took an interim test ($M = 73.33$, $SD = 14.68$) scored significantly higher on the final test of Section A than those who were not tested ($M = 67.67$, $SD = 14.62$), demonstrating a backward effect of testing, $d = 0.39$. Furthermore, there was a significant main effect of feature highlighting, $F(1, 96) = 7.91$, $p = 0.006$, $\eta_p^2 = 0.076$, $BF_{incl} = 7.05$, indicating moderate evidence for the inclusion of feature highlighting in the model. The final test score on Section A was significantly worse when feature highlighting was provided ($M = 66.50$, $SD = 14.58$) than when it was absent ($M = 74.50$, $SD = 14.15$), $d = 0.56$. There was no significant interaction between interim activity and feature highlighting, $F(1, 96) = 0.01$, $p = 0.907$, $BF_{excl} = 2.74$. There were no other significant two-way or three-way interactions.

To investigate the forward effect of testing, we examined the performance for Section B. There was a significant main effect of test item type, $F(1, 96) = 39.53$, $p < 0.001$, $\eta_p^2 = 0.292$, $BF_{incl} = 686{,}499.28$, indicating decisive evidence for the inclusion of the test item type in the model. The percentage of correctly classified rock images was significantly higher for the old items ($M = 72.08$, $SD = 21.07$) than for the new items ($M = 63.75$, $SD = 18.17$), $d = 0.62$. There was a significant main effect of interim activity, $F(1, 96) = 4.09$, $p = 0.046$, $\eta_p^2 = 0.041$, $BF_{incl} = 1.54$, indicating anecdotal evidence for the inclusion of the interim activity in the model. The participants who took an interim test ($M = 71.42$, $SD = 17.19$) scored significantly higher on the final test of Section B than those who restudied ($M = 64.42$, $SD = 19.24$), demonstrating a forward effect of testing, $d = 0.38$. Moreover, there was a significant main effect of feature highlighting, $F(1, 96) = 12.68$, $p < 0.001$, $\eta_p^2 = 0.117$, $BF_{incl} = 47.05$, indicating very strong evidence for the inclusion of feature highlighting in the model. The final test score on Section B was significantly worse when feature highlighting was provided ($M = 61.75$, $SD = 18.69$) than when

it was absent ($M = 74.08$, $SD = 16.24$), $d = 0.71$. There was no significant interaction between interim activity and feature highlighting, $F(1, 96) = 0.06$, $p = 0.810$, $BF_{excl} = 2.05$. The only significant interaction was found between test item type and interim activity, $F(1, 96) = 4.57$, $p = 0.035$, $\eta_p^2 = 0.045$, $BF_{incl} = 1.62$, indicating anecdotal evidence for the inclusion of the interaction in the model. This was because the forward effect of testing was significant for the old items, $t(98) = 2.39$, $p = 0.019$, $d = 0.48$, but not for the new items, $t(98) = 1.15$, $p = 0.254$, implying that the forward effect of testing was more apparent for retention than for transfer. There were no other significant two-way or three-way interactions.

Discussion

Experiment 1 investigated the effects of interim testing and feature highlighting on natural category learning. First, the results revealed that participants performed better on the old items than on the new ones, suggesting that the retention test was relatively easier than the transfer test. Bayesian analysis indicated decisive evidence for the inclusion of the test item type for both Sections A and B. This is not surprising given that old items were presented to participants multiple times whereas new items were presented for the first time in the final test.

Second, interim testing facilitated natural category learning. Participants who took an interim test outperformed those who restudied the materials from both Sections A and B on the final test. That is, interim testing on previously studied materials enhanced not only the learning of that tested categories (Section A), indicating the backward effect of testing, but also the learning of subsequently studied categories (Section B), indicating the forward effect of testing. The results are consistent with previous work on category learning that revealed the backward effect (Jacoby et al., 2010) and forward effect (Choi & Lee, 2020; Lee & Ahn, 2018; Lee & Ha, 2019; Yang & Shanks, 2018) of testing. However, Bayesian analyses indicated only anecdotal evidence for the inclusion of interim activity for both Sections A and B. This might be because the format of the interim test was not conducive to the final test format, especially for the FH test condition, and we will discuss this issue further in the following paragraph. More importantly, there was no significant interaction between interim activity and feature highlighting. The only significant interaction was identified between test item type and interim activity for Section B, suggesting that the forward effect of testing was more apparent on retention items than on transfer items.

Surprisingly, feature highlighting did not further enhance learning but had a negative impact. Participants who were not given feature highlighting outperformed those who were provided during the entire learning and interim test sessions. Bayesian analyses revealed moderate evidence for Section A and very strong evidence for Section B for the inclusion of feature highlighting. The results are inconsistent with the previous work (Miyatsu et al., 2019) that showed positive effects of feature highlighting on natural category learning. One possible explanation is that the participants who were given feature highlighting developed test format expectations not conducive to the final test and accordingly employed non-optimal study strategies. In the FH conditions, feature highlighting was included in both test problem images and feedback during the interim

test, thus participants might have expected that they would be given feature highlighting in a later final test as well. If this was the case, participants might have read and memorized explicit verbal descriptions, rather than processing perceptual features of the rock images. As a result, they perhaps had difficulty retrieving the corresponding category name when they were not given feature highlighting on the problem images of the final test. Conversely, in the no-FH conditions, participants were never presented with any explicit instructions on the features of rock categories. They probably developed a test format expectation more consistent with the format of the final test. Indeed, several previous studies have shown that students adopt different study strategies depending on their test format expectations (e.g., Abel & Bäuml, 2016; Finley & Benjamin, 2012; Middlebrooks et al., 2017). Experiment 2 explored these possibilities by eliminating feature highlighting from the interim test problem images.

Experiment 2

In Experiment 2, we removed feature highlighting from the test problem images but presented it only as feedback during the interim test session; thus, the participants in the FH test condition were presented with the same interim test problem images (i.e., they had to classify rock images without feature highlighting) as in the no-FH test condition.

Method

Participants

Participants were recruited from a large university in South Korea via school-wide online advertisements. A total of 121 undergraduates participated online in return for monetary compensation. As in Experiment 1, we excluded three participants whose test scores were above or below the three *SD* of the group mean, 11 participants who reported a high level of prior knowledge in rock categories, and seven participants who left the experiment idle on the self-paced page for longer than 10 min. Thus, there were 100 participants (71 women; 29 men; mean age = 21.84 years) for the final analyses. Each participant was randomly assigned to one of four conditions: 25 in the FH test, 25 in the no-FH test, 25 in the FH restudy, and 25 in the no-FH restudy.

Design, Materials, and Procedure

The design and procedure in Experiment 2 were identical to those in Experiment 1, except that in the FH test condition feature highlighting was not provided in the test problem images but it was provided only as feedback during the interim test session. In the interim test session, participants were given the rock images without feature highlighting, and they were able to see feature highlighting only after entering the name of the rock category for each image. Thus, the format of the interim test in both the FH and no-FH conditions was the same as that of the final test.

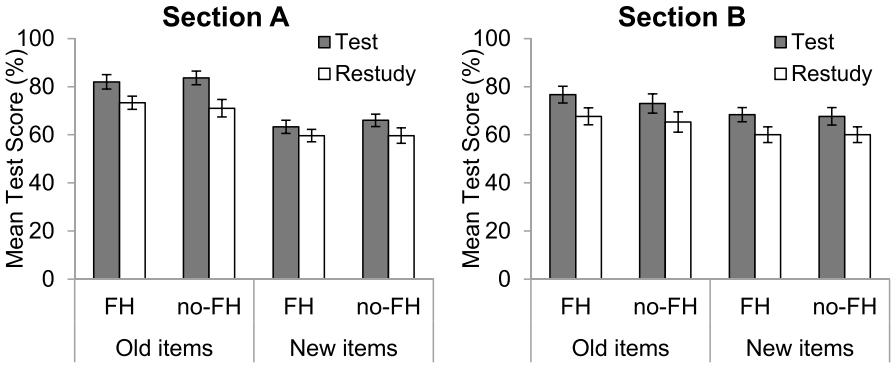

Fig. 3 Mean percentage of the correct responses for the old and new items in the final test of Sections A and B in the FH test, no-FH test, FH restudy, and no-FH restudy conditions of Experiment 2. Error bars represent one standard error of the mean

Results

Interim Test Performance

As in Experiment 1, only the participants in the FH and no-FH test conditions were tested on Section A. The mean interim test score was 78.45 ($SD = 14.51$) in the FH test condition and 79.78 ($SD = 14.22$) in the no-FH test condition. The mean difference between the FH and no-FH test conditions was not significant, $t(48) = 0.33$, $p = 0.744$.

Final Test Performance

Figure 3 shows the mean percentage of correct classification for the old (retention) and new rock images (transfer) in the final test of Sections A and B. Two 2×2×2 mixed ANOVAs were conducted to investigate the backward and forward effects of testing on the percentage of correctly classified rock images, one on each for Sections A and B. As in Experiment 1, feature highlighting (FH vs. no-FH) and interim activity (test vs. restudy) were included as between-subjects factors, while the test item type (old vs. new) was included as a within-subject factor.

For Section A, there was a significant main effect of test item type, $F(1, 96) = 128.61$, $p < 0.001$, $\eta_p^2 = 0.573$, $BF_{incl} = 2.29 \times 10^{+16}$, indicating decisive evidence for the inclusion of the test item type in the model. The percentage of correctly classified rock images was significantly higher for the old items ($M = 77.50$, $SD = 16.09$) than for the new items ($M = 62.17$, $SD = 14.09$), $d = 1.12$. As in Experiment 1, there was a significant main effect of the interim activity, $F(1, 96) = 8.97$, $p = 0.003$, $\eta_p^2 = 0.085$, $BF_{incl} = 11.53$, indicating strong evidence for the inclusion of the interim activity in the model. The participants who took an interim test ($M = 73.75$, $SD = 11.91$) scored significantly higher on the final test of Section A than those who were not tested ($M = 65.92$, $SD = 13.95$), $d = 0.60$, demonstrating a backward effect of testing. More importantly, unlike in Experiment 1, feature

highlighting did not have a significant effect, $F(1, 96) = 0.04$, $p = 0.849$, $BF_{excl} = 3.44$, indicating moderate evidence for the exclusion of feature highlighting in the model. The final test performance on Section A was not significantly different between the FH ($M = 70.08$, $SD = 14.25$) and no-FH conditions ($M = 69.58$, $SD = 12.83$). Furthermore, there was no significant interaction between the interim activity and feature highlighting, $F(1, 96) = 0.41$, $p = 0.525$, $BF_{excl} = 2.34$. The only significant interaction was test item type by interim activity interaction, $F(1, 96) = 4.39$, $p = 0.039$, $\eta_p^2 = 0.044$, $BF_{incl} = 1.41$, indicating anecdotal evidence for the inclusion of the interaction in the model. This was because the effect of testing was significant for the old items, $t(98) = 3.50$, $p < 0.001$, $d = 0.70$, but not for the new items, $t(98) = 1.80$, $p = 0.076$, implying that the backward testing effect was more apparent for retention than for transfer. There were no other significant two-way or three-way interactions.

For Section B, there was a significant main effect of test item type, $F(1, 96) = 25.25$, $p < 0.001$, $\eta_p^2 = 0.208$, $BF_{incl} = 7761.70$, indicating decisive evidence for the inclusion of the test item type in the model. The percentage of correctly classified examples was significantly higher for the old items ($M = 70.67$, $SD = 19.37$) than for the new items ($M = 64.00$, $SD = 16.66$), $d = 0.51$. Consistent with Experiment 1, there was a significant main effect of interim activity, $F(1, 96) = 6.09$, $p = 0.015$, $\eta_p^2 = 0.060$, $BF_{incl} = 3.51$, indicating moderate evidence for the inclusion of the interim activity in the model. The participants who took an interim test ($M = 71.42$, $SD = 15.97$) scored significantly higher on the final test of Section B than those who restudied ($M = 63.25$, $SD = 16.83$), $d = 0.50$, demonstrating a forward effect of testing. Moreover, unlike in Experiment 1, feature highlighting did not have a significant effect, $F(1, 96) = 0.25$, $p = 0.616$, $BF_{excl} = 2.50$, indicating anecdotal evidence for the exclusion of feature highlighting in the model. The final test performance on Section B was not significantly different between the FH ($M = 66.50$, $SD = 18.25$) and no-FH conditions ($M = 68.17$, $SD = 15.43$). Furthermore, there was no significant interaction between interim activity and feature highlighting, $F(1, 96) = 0.02$, $p = 0.880$, $BF_{excl} = 2.26$. No other two-way or three-way interactions were significant.

Discussion

Replicating the results from Experiment 1, there was a robust effect of test item type such that participants performed better for old items than new items. Bayesian analyses revealed decisive evidence for the inclusion of the test item type for both Sections A and B. More importantly, Experiment 2 again showed the benefits of testing. Participants who took an interim test on Section A outperformed those who restudied in both Sections A and B, indicating both the backward and forward effects of testing in natural category learning. Also, the positive effect of testing was larger in Experiment 2 than in Experiment 1. Bayesian analyses indicated strong evidence for Section A and moderate evidence for Section B for the inclusion of interim activity in the model. This was probably because of the changed format of the interim test. In Experiment 2,

we removed feature highlighting from the test problem images and presented it only as feedback. The removal of feature highlighting perhaps encouraged participants to put more effort into retrieval during interim testing, and such increased retrieval effort in turn increased testing effects (Pyc & Rawson, 2009). The changed format of the interim test may also have allowed participants to expect the correct format of the final test and thus probably better prepared them for the final test. Furthermore, a significant interaction between test item type and interim activity was found for Section A, which was also observed for Section B in Experiment 1, implying that the testing effects may be more apparent in retention than in transfer.

More importantly, the negative effect of feature highlighting shown in Experiment 1 disappeared in Experiment 2. Participants in the FH conditions showed as good performance as those in the no-FH conditions. This was probably due to the control of test format expectations by making the test format of the interim and final tests identical. However, we again did not obtain any benefits of feature highlighting in natural category learning. The groups that were given feature highlighting never outperformed those that were not. Bayesian analyses revealed that the data were about 3.44 (for Section A) and 2.50 (for Section B) times more likely in models that did not include the feature highlighting variable than those that included. Also, feature highlighting did not interact with interim activity, suggesting that the inclusion of explicit rules in learning materials may not increase testing benefits.

One possible explanation for the lack of feature highlighting benefit is that feature highlighting in the learning session had students heavily rely on verbal descriptions and distract them from studying the rock images. When feature highlighting was presented, participants had to spend their viewing time reading the verbal descriptions, which could compete with the viewing time of the rock images. Especially during the initial learning session, it may be more effective to process visual images rather than reading verbal descriptions in order to identify key perceptual features. Indeed, a previous study that demonstrated the benefit of feature highlighting (Miyatsu et al., 2019) used feature highlighting only as feedback by employing a feedback training method. In addition, the null effect of feature highlighting in Experiment 2 could be because participants did not have sufficient time to process feedback. It is possible that students did not fully comprehend the feature descriptions during the interim test session because feature highlighting was only offered for 2 s as feedback. Another limitation of Experiments 1 and 2 was that the sample size was relatively small compared to the previous research. In Miyatsu et al.'s (2019, Experiment 2) study, there were a total of 80 participants when comparing the FH present versus FH absent between-subjects conditions. It is possible that the effect of feature highlighting was not significant because the sample size in Experiments 1 and 2 was insufficient to detect small effects. Thus, Experiment 3 was conducted with several modifications to address these possibilities.

Experiment 3

Experiment 3[1] aimed to investigate whether the null effect of feature highlighting would be consistent with several modifications. First, we did not provide feature highlighting during the learning session of Sections A and B but presented it only as feedback during the interim test session. Second, we increased the duration of feedback time. Third, we increased the sample size to 60 participants per condition. Finally, we recruited participants from all over the country using online panels in order to improve the generalizability of our findings.

Method

Participants

We recruited participants using dataSpring (https://www.d8aspring.com/) online panels and participation was restricted to undergraduate students from South Korea. To detect a small ($\eta_p^2 = 0.04$) to medium-sized effect ($\eta_p^2 = 0.06$) of feature highlighting at a power of 0.9, the required sample size was between 42 and 64 per condition. Thus, we recruited participants until reaching 60 per condition after excluding some participants based on the exclusion criteria. A total of 299 undergraduates participated online in exchange for monetary compensation. We excluded six participants who repeatedly submitted the same response, 24 participants who reported taking notes, 13 participants who reported a high level of prior knowledge in rock categories, and 16 participants who left the experiment idle on the self-paced page for longer than 10 min. Thus, there were 240 participants (157 women; 83 men; mean age = 21.53 years) for the final analyses. Each participant was randomly assigned to one of four conditions: 60 in the FH test, 60 in the no-FH test, 60 in the FH restudy, and 60 in the no-FH restudy.

Design, Materials, and Procedure

The design and procedure of Experiment 3 were identical to those of Experiments 1 and 2, except for the two changes. First, feature highlighting was removed from the entire learning session of both Sections A and B and the interim test problem images. In the FH test condition, feature highlighting was included only in the feedback images. This procedure was in line with the feedback training approach, in which previous research demonstrated the benefits of feature highlighting (Miyatsu

[1] Prior to Experiment 3, we ran an additional experiment with the two test conditions (FH and no-FH) with feature highlighting only provided as feedback during the interim test session for the FH-test condition as in Experiment 3. The results consistently showed the null effect of feature highlighting. However, due to several critical methodological limitations (e.g., short feedback duration, small sample size, lack of restudy conditions), we decided to run Experiment 3 by addressing all of these limitations. The data of this unreported experiment are available at the following link with all other experimental data: https://osf.io/mt2e9/?view_only=753be8b7f23f4931977c4b35db6b266a.

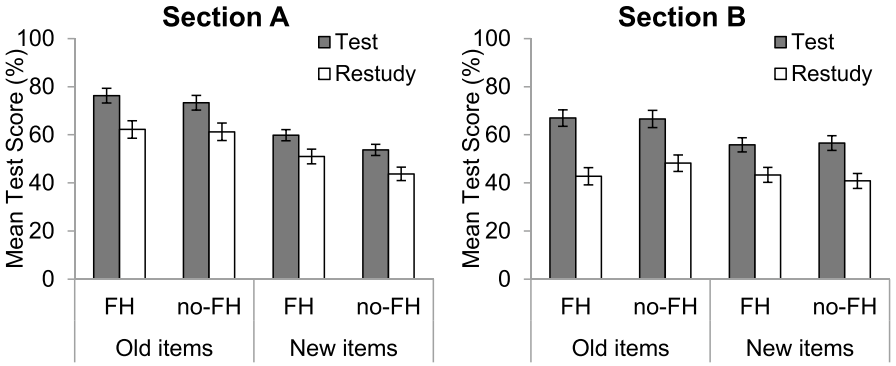

Fig. 4 Mean percentage of the correct responses for the old and new items in the final test of Sections A and B in the FH test, no-FH test, FH restudy, and no-FH restudy conditions of Experiment 3. Error bars represent one standard error of the mean

et al., 2019). In the FH restudy condition, feature highlighting was included only in the rock images presented during the interim restudy session. Thus, feature highlighting appeared only during the interim activity, either as feedback after each interim test trial for the FH test condition, or as restudy material for the FH restudy condition. The no-FH test and no-FH restudy conditions were the same as in Experiments 1 and 2. Second, feedback was presented for 4 s during the interim test session. In order to equalize the interim session time, participants in the restudy conditions restudied each rock image for 8 s.

Results

Interim Test Performance

As in Experiments 1 and 2, only the participants in the FH and no-FH test conditions were tested on Section A. The mean interim test score was 69.63 ($SD = 24.87$) in the FH test condition and 72.13 ($SD = 21.05$) in the no-FH test condition. The mean difference between the FH and no-FH test conditions was not significant, $t(118) = 0.59$, $p = 0.553$.

Final Test Performance

Figure 4 shows the mean percentage of correct classification for the old (retention) and new rock images (transfer) in the final test of Sections A and B. Two separate $2 \times 2 \times 2$ mixed ANOVAs were conducted to investigate the backward and forward effects of testing on the percentage of correctly classified rock images for Sections A and B. As in Experiments 1 and 2, feature highlighting (FH vs. no-FH) and interim activity (test vs. restudy) were included as between-subjects factors, while the test item type (old vs. new) was included as a within-subject factor.

For Section A, there was a significant main effect of test item type, $F(1, 236) = 258.32$, $p < 0.001$, $\eta_p^2 = 0.523$, $BF_{incl} = 4.55 \times 10^{+35}$, indicating decisive evidence for the inclusion of the test item type in the model. The percentage of correctly classified rock images was significantly higher for the old items ($M = 68.26$, $SD = 26.69$) than for the new items ($M = 52.08$, $SD = 21.04$), $d = 1.02$. There was a significant main effect of interim activity, $F(1, 236) = 15.67$, $p < 0.001$, $\eta_p^2 = 0.062$, $BF_{incl} = 209.85$, indicating decisive evidence for the inclusion of the interim activity in the model. The participants who took an interim test ($M = 65.80$, $SD = 19.45$) scored significantly higher on the final test of Section A than those who were not tested ($M = 54.55$, $SD = 24.33$), $d = 0.51$, demonstrating a backward effect of testing. More importantly, feature highlighting did not have a significant effect, $F(1, 236) = 2.30$, $p = 0.131$, $BF_{excl} = 1.44$, indicating anecdotal evidence for the exclusion of feature highlighting in the model. The final test performance on Section A was not significantly different between the FH ($M = 62.33$, $SD = 23.04$) and no-FH conditions ($M = 58.02$, $SD = 22.23$). There was no significant interaction between the interim activity and feature highlighting, $F(1, 236) = 0.01$, $p = 0.942$, $BF_{excl} = 2.99$. The only significant interaction was test item type by feature highlighting interaction, $F(1, 236) = 5.50$, $p = 0.020$, $\eta_p^2 = 0.023$, $BF_{incl} = 1.88$. This was because the performance difference between the old and new items was greater for the no-FH groups (mean difference = 18.54, $p < 0.001$, $d = 1.12$) than for the FH groups (mean difference = 13.82, $p < 0.001$, $d = 0.94$), although both groups showed better classification performance for the old than the new items. There were no other significant two-way or three-way interactions.

For Section B, there was a significant main effect of test item type, $F(1, 236) = 50.23$, $p < 0.001$, $\eta_p^2 = 0.176$, $BF_{incl} = 1.04 \times 10^{+8}$, indicating decisive evidence for the inclusion of the test item type in the model. The percentage of correctly classified examples was significantly higher for the old items ($M = 56.11$, $SD = 29.10$) than for the new items ($M = 49.13$, $SD = 24.49$), $d = 0.44$. There was a significant main effect of interim activity, $F(1, 236) = 31.80$, $p < 0.001$, $\eta_p^2 = 0.119$, $BF_{incl} = 221{,}149.93$, indicating decisive evidence for the inclusion of the interim activity in the model. The participants who took an interim test ($M = 61.46$, $SD = 24.09$) scored significantly higher on the final test of Section B than those who restudied ($M = 43.79$, $SD = 24.27$), $d = 0.73$, demonstrating a forward effect of testing. However, feature highlighting did not have a significant effect, $F(1, 236) = 0.07$, $p = 0.799$, $BF_{excl} = 3.21$, indicating moderate evidence for the exclusion of feature highlighting in the model. The final test performance on Section B was not significantly different between the FH ($M = 52.22$, $SD = 25.70$) and no-FH conditions ($M = 53.02$, $SD = 25.81$). There was no significant interaction between interim activity and feature highlighting, $F(1, 236) = 0.04$, $p = 0.833$, $BF_{excl} = 2.65$. The test item type by interim activity interaction was significant, $F(1, 236) = 13.19$, $p < 0.001$, $\eta_p^2 = 0.053$, $BF_{incl} = 52.79$. This was because the performance difference between the old and new items was greater for the interim test groups (mean difference = 10.56, $p < 0.001$, $d = 0.72$) than for the interim restudy groups (mean difference = 3.40, $p = 0.024$, $d = 0.21$), although both groups showed better classification performance for the old than the new items. The test item type * feature highlighting * interim activity interaction was also significant, $F(1, 236) = 5.25$, $p = 0.023$, $\eta_p^2 = 0.022$,

$BF_{incl} = 1.99$. This was because the test item type by interim activity interaction was significant for the FH conditions, $F(1, 118) = 16.40$, $p < 0.001$, $\eta_p^2 = 0.122$, but not for the no-FH conditions $F(1, 118) = 0.97$, $p = 0.328$. The test item type by feature highlighting interaction was not significant, $F(1, 236) = 2.99$, $p = 0.085$.

Discussion

Experiment 3 examined the effect of feature highlighting and interim testing with a larger sample and replicated the main findings of Experiments 1 and 2. First, participants performed better for old items than for new items. Bayesian analyses again revealed decisive evidence for the inclusion of the test item type for both Sections A and B, consistent with the results of Experiments 1 and 2. Second, participants who took an interim test on Section A outperformed those who restudied items in both Sections A and B, indicating both the backward and forward effects of testing in natural category learning. Also, we observed stronger evidence for the benefits of interim testing in Experiment 3 than in Experiments 1 and 2. Bayesian analyses indicated decisive evidence for both Sections A and B for the inclusion of interim activity. This is probably because of the increased sample size. Third, most importantly, we again failed to obtain any positive effects of feature highlighting, even with a larger sample. Bayesian analyses revealed that the data were about 1.44 (for Section A) and 3.21 (for Section B) times more likely in models that did not include the feature highlighting variable than those that included. Further, feature highlighting did not interact with interim activity, suggesting that the inclusion of explicit rules in learning materials did not increase testing benefits. In Experiment 3, the feedback presentation time was increased to 4 s to provide sufficient time to process verbal descriptions, but it did not seem to enhance learning. These results again conflict with those of a previous study that demonstrated the positive effect of feature highlighting in natural category learning (Miyatsu et al., 2019). We will discuss possible explanations for these findings in General Discussion.

General Discussion

The present study aimed to investigate whether incorporating feature highlighting in learning materials could result in greater testing benefits by examining both the backward and forward effects of testing in natural category learning. We had participants study 12 different rock categories divided into two learning sections (Sections A & B) and asked them to either take an interim test or not between the two learning sections. The interim test was only performed on Section A, and all participants studied Section B under the same conditions. This procedure allowed us to distinguish the effect of interim testing on Section A (the backward effect of testing) and that on Section B (the forward effect of testing). Additionally, the presence of feature highlighting was manipulated during the learning and interim test sessions to examine an effective way of providing feature highlighting. More specifically,

feature highlighting was provided during the entire learning and interim test sessions (Experiment 1), during the learning session and as feedback during the interim test session (Experiment 2), or only as feedback during the interim test session (Experiment 3). This manipulation allowed us to examine how and when feature highlighting should be given to increase testing benefits.

The results showed both the backward and forward effects of testing in natural category learning. Taking an interim test on studied categories enhanced both the learning of those categories (Section A) and the learning of subsequently studied new categories (Section B). This is consistent with previous studies that demonstrated the backward effect (Jacoby et al., 2010) and the forward effect (Choi & Lee, 2020; Lee & Ahn, 2018; Lee & Ha, 2019; Yang & Shanks, 2018; Yang et al., 2019) of testing in category learning. The finding is also consistent with the prior studies that compared the classification performance between feedback training and observational training (e.g., Ashby et al., 2002; Edmunds et al., 2015) in natural category learning. Considering that our test condition is more similar to feedback training, we showed that feedback training was more effective than observational training. Our findings contribute to a large body of research on the broad application of testing by showing the benefits of testing in natural category learning.

To compute the mean effect size of testing, we conducted a mini meta-analysis integrating the results from all three experiments. A random effects model revealed significant testing effects for both old (retention) and new (transfer) items in both Sections A and B. In Section A (backward testing effect), the testing effect was larger for old items, $d=0.54$, 95% CI [0.35 0.73], than for new items, $d=0.38$, 95% CI [0.19, 0.57]. In Section B (forward testing effect), the testing effect was again larger for old items, $d=0.61$, 95% CI [0.36, 0.85], than for new items, $d=0.48$, 95% CI [0.26, 0.70]. Notably, testing was more beneficial for old items than for new items, both from the studied categories of Sections A and B, indicating that both backward and forward benefits of testing were larger on the retention test than on the transfer test. This finding is consistent with a study by Jacoby et al. (2010), which showed that testing was more beneficial for studied items than for unstudied items of the previously studied categories (backward testing effect). The current study extends this finding by demonstrating that testing was also more effective for studied items than unstudied items from the newly studied categories after taking a test (forward testing effect). The observed larger benefit for studied (vs. unstudied) items could be because the final test was taken immediately after learning. Several studies have found that testing has a stronger effect after a delay (Carpenter, 2012). Another possible explanation is the difference in exposure between old and new items. Participants were exposed to the old items twice before the final test, while the new items were presented for the first time in the final test. Thus, for the old items in the final test, the test groups perhaps may have only needed to retrieve the old items that they had already practiced during the interim test. In contrast, for the new items, participants had to generalize their learning, which they had not practiced during the interim test. The testing effects could be dependent on how much the processing that occurred during the interim test corresponds to the processing that occurred during the final test (transfer-appropriate processing, Morris et al., 1977).

What we repeatedly found across the three experiments is that testing enhanced category learning. The observed backward testing effect on category learning could be explained by active retrieval. Testing allows students to learn more actively by retrieving information from their memory (Bjork & Bjork, 2011). In the present study, the increased testing effects as experiments progressed lend support to this account, as Bayesian analyses revealed only anecdotal evidence for Experiment 1; strong (Section A) and moderate (Section B) evidence for Experiment 2; and decisive evidence for Experiment 3. One of the important changes across the three experiments was the reduction of the amount of feature highlighting. This probably increased the learners' retrieval effort in the test conditions, increasing the benefits of testing. Consistent with this explanation, previous studies have shown that more effortful retrieval leads to greater learning performance (Butler & Roediger, 2007; McDaniel et al., 2007). Further, the removal of feature highlighting may also have strengthened the forward testing effect by increasing retrieval effort as proposed in previous studies (Cho et al., 2017; Yang et al., 2019). If participants in the FH condition found it easier to retrieve the category names with feature highlighting during the interim test, they could have put less cognitive effort in subsequent learning. Additionally, it is possible that participants expected feature descriptions to be provided during the final test, which may have led them to rely heavily on verbal cues rather than processing other visual cues in the rock images in later learning. Future research will need to investigate how different levels of retrieval effort influence the effect of testing on category learning.

Different from our hypothesis, however, this study found no evidence for a beneficial effect of feature highlighting on natural category learning. Although removing feature highlighting from the interim test materials eliminated the negative effect of feature highlighting found in Experiment 1, a positive effect was never demonstrated in Experiments 2 and 3. Even when the viewing time of feature highlighting feedback was increased to 4 s (Experiment 3), as in previous work (Miyatsu et al., 2019), this did not improve learning. Indeed, Bayesian analyses revealed anecdotal to moderate evidence in favor of the null effect of feature highlighting in both Experiments 2 and 3. Also, the mean effect size computed from a random-effects meta-analysis of all three experiments was not significant for Section A, $d = -0.09$, 95% CI [-0.53, 0.35], and Section B, $d = -0.26$, 95% CI [-0.66, 0.15]. Our results were consistent with those of more recent research (Whitehead et al., 2021), which also did not observe the benefits of feature highlighting.

Several category learning models have proposed that during the initial category learning, learners generate simple and explicit verbal rules (e.g., Ashby et al., 1998; Nosofsky et al., 1994). Accordingly, we hypothesized that students would use feature highlighting as the rules, which would reduce the cognitive burden for inefficient hypothesis search and testing and increase the benefits of testing. In contrast to our hypothesis, the interim activity did not interact with feature highlighting in any of the reported experiments. Additionally, Bayesian analyses revealed anecdotal evidence for the null effect of interim activity by feature highlighting interaction. When we increased the sample size in Experiment 3, significant testing effects were obtained regardless of the presence of feature highlighting; feature highlighting neither increased nor decreased the testing benefits. There was no statistically

significant difference between the FH test and no-FH test conditions in either the interim or final test scores.

One possible explanation for the lack of feature highlighting benefit is the limited viewing time given during the learning session. Previous research has raised concerns about the insufficient viewing time required to process rock images, which might compete with the viewing time needed to read verbal descriptions (Meagher et al., 2022; Miyatsu et al., 2019). In the present study, participants were given 4 s for each rock image during the learning sessions and 2 s (Experiments 1 and 2) or 4 s (Experiment 3) for each feedback page during the interim test session where they had to visually process both textual and pictorial information in an experimenter-paced manner. Although we determined the duration of each trial based on our pilot and previous studies (Lee & Ahn, 2018; Miyatsu et al., 2019), some participants might still find it short. The fact that the negative effect of feature highlighting disappeared when the feature highlighting was removed from the rock exemplars in learning sessions (Experiment 3) and interim test problem images (Experiments 2 and 3) also suggests the plausibility of this account. When feature highlighting was eliminated, the FH groups may have had sufficient time to process the entire rock image without being distracted by verbal descriptions. Also, processing feature highlighting may not necessarily reduce cognitive burden as we hypothesized. Although the provision of feature highlighting can reduce inefficient hypothesis search and testing, it can instead increase the cognitive burden for processing both textual and pictorial information. Additionally, one notable difference between Miyatsu et al.'s (2019) and our research was the duration of the entire learning session. In the previous study, participants observed each exemplar once during the observation phase and 24 times (2 repetitions × 12 blocks for each rock exemplar) during the feedback learning phase. In the current study, participants observed each exemplar once during Section A and once during the interim session. The shorter learning session in the current study may have made it difficult to obtain the benefits of feature highlighting. The effectiveness of learning methods can vary depending on the amount of practice (e.g., Brunstein et al., 2009). The benefits of feature highlighting may be limited when learners do not have sufficient time to integrate the information into their understanding. Also, the effect of feature highlighting may change depending on the characteristics of learning materials (e.g., Rosedahl et al., 2021). Given mixed findings regarding the effect of feature highlighting, future research should examine whether the effect of the feature highlighting changes depending on the learning context (e.g., self- vs. experimenter-paced study) and type of learning materials.

Furthermore, our study did not reveal the positive effect of combining two learning strategies (interim testing and feature highlighting) on natural category learning. Many prior studies have shown that combining multiple empirically supported strategies do not always lead to learning enhancement. For example, Whitehead et al. (2021) employed feature highlighting with interleaving and did not find a positive effect of feature highlighting. Given that recent studies and ours have failed to show the benefits of using feature highlighting with another empirically supported strategy, future research should investigate when feature highlighting does and does not enhance natural category learning and how it can be successfully combined with other instructional methods.

Conclusion and Practical Implications

The findings of the present study suggest that providing more information does not necessarily facilitate natural category learning in two ways. First, giving information (restudy) was less effective than withholding information (testing). When taking a test, learners must retrieve information from their memory without being presented with the learning materials. However, when restudying, learners are provided with all the information. Our results indicate that active retrieval attempts are more beneficial for natural category learning than passive restudying. The present study also showed that the forward effect of testing can be expanded to natural category learning, in addition to replicating prior work on the backward effect of testing (Jacoby et al., 2010). With numerous studies on the testing effects (McDermott, 2021; Rowland, 2014), the present study demonstrates that testing is an effective learning tool for various learning tasks. Second, feature highlighting did not further enhance learning. These findings again suggest that providing additional information is not necessarily more beneficial. In the present study, participants in the FH conditions could use the given feature highlighting to distinguish categories, while those in the no-FH conditions had to learn categories by abstracting the features of each category themselves. Although less information was provided, participants in the no-FH conditions performed better or showed as good performance as those in the FH conditions. Feature highlighting was especially harmful when presented more (Experiment 1) than less (Experiments 2 and 3) during learning, further indicating that more information is not necessarily better than less information.

This study has important educational implications. Given that our study showed both the backward and forward effects of testing in natural category learning, science teachers may be able to improve students' category learning by interpolating tests between learning segments and encouraging students to use self-testing. However, the use of feature highlighting in natural category learning requires some caution. Even when textual descriptions are explicitly provided with pictorial examples in many educational resources, the addition of such verbal information may not guarantee more effective learning. It may even hinder learning, especially when the processing of textual information interferes with the processing of pictorial information.

Data Availability All data have been made publicly available at the Open Science Framework (OSF) and can be accessed at https://osf.io/mt2e9/?view_only=753be8b7f23f4931977c4b35db6b266a.

Declarations

Preliminary results were presented at the meeting of the Cognitive Science Society, 2021.

Ethics Approval All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional research committee and with the 1964 Helsinki Declaration and its later amendments or comparable ethical standards. Approval was obtained from the Institutional Review Board of Yonsei University.

Consent to Participate Informed consent was obtained from all individual participants included in the study.

Competing Interests The authors have no relevant financial or non-financial interests to disclose.

References

Abel, M., & Bäuml, K. H. T. (2016). Retrieval practice can eliminate list method directed forgetting. *Memory & Cognition, 44*(1), 15–23. https://doi.org/10.3758/s13421-015-0539-x

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. https://doi.org/10.1037/0033-295X.105.3.442

Ashby, F. G., Maddox, W. T., & Bohil, C. J. (2002). Observational versus feedback training in rule-based and information-integration category learning. *Memory & Cognition, 30*, 666–677. https://doi.org/10.3758/BF03196423

Ashby, F. G., Queller, S., & Berretty, P. M. (1999). On the dominance of unidimensional rules in unsupervised categorization. *Perception & Psychophysics, 61*, 1178–1199. https://doi.org/10.3758/BF03207622

Birnbaum, M. S., Kornell, N., Bjork, E. L., & Bjork, R. A. (2013). Why interleaving enhances inductive learning: The roles of discrimination and retrieval. *Memory & Cognition, 41*(3), 392–402. https://doi.org/10.3758/s13421-012-0272-7

Bjork, E. L., & Bjork, R. A. (2011). Making things hard on yourself, but in a good way: Creating desirable difficulties to enhance learning. In M. A. Gernsbacher, R. W. Pew, L. M. Hough, & J. R. Pomerantz (Eds.), *Psychology and the real world: Essays illustrating fundamental contributions to society* (pp. 56–64). Worth Publishers.

Brunstein, A., Betts, S., & Anderson, J. R. (2009). Practice enables successful learning under minimal guidance. *Journal of Educational Psychology, 101*(4), 790–802. https://doi.org/10.1037/a0016656

Butler, A. C., & Roediger, H. L. (2007). Testing improves long-term retention in a simulated classroom setting. *European Journal of Cognitive Psychology, 19*(4–5), 514–527. https://doi.org/10.1080/09541440701326097

Carpenter, S. K. (2012). Testing enhances the transfer of learning. *Current Directions in Psychological Science*, *21*(5), 279–283. https://doi.org/10.1177/0963721412452728

Carvalho, P. F., & Goldstone, R. L. (2015). The benefits of interleaved and blocked study: Different tasks benefit from different schedules of study. *Psychonomic Bulletin & Review, 22*(1), 281–288. https://doi.org/10.3758/s13423-014-0676-4

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

Chan, J. C. K., Meissner, C. A., & Davis, S. D. (2018b). Retrieval potentiates new learning: a theoretical and meta-analytic review. *Psychological Bulletin, 144*(11), 1111–1146. https://doi.org/10.1037/bul0000166

Cho, K. W., Neely, J. H., Crocco, S., & Vitrano, D. (2017). Testing enhances both encoding and retrieval for both tested and untested items. *The Quarterly Journal of Experimental Psychology, 70*(7), 1211–1235. https://doi.org/10.1080/17470218.2016.1175485

Choi, H., & Lee, H. S. (2020). Knowing is not half the battle: The role of actual test experience in the forward testing effect. *Educational Psychology Review, 32*(3), 765–789. https://doi.org/10.1007/s10648-020-09518-0

Cummings, E. L., Reeb, A., & McDaniel, M. A. (2022). Do not forget the keyword method: Learning educational content with arbitrary associations. *Journal of Applied Research in Memory and Cognition, 12*(1), 70–81. https://doi.org/10.1037/mac0000031

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

Edmunds, C. E. R., Milton, F., & Wills, A. J. (2015). Feedback can be superior to observational training for both rule-based and information-integration category structures. *Quarterly Journal of Experimental Psychology, 68*(6), 1203–1222. https://doi.org/10.1080/17470218.2014.978875

Eglington, L. G., & Kang, S. H. K. (2017). Interleaved presentation benefits science category learning. *Journal of Applied Research in Memory and Cognition, 6*(4), 475–485. https://doi.org/10.1016/j.jarmac.2017.07.005

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

Finley, J. R., & Benjamin, A. S. (2012). Adaptive and qualitative changes in encoding strategy with experience: Evidence from the test-expectancy paradigm. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 38*(3), 632–652. https://doi.org/10.1037/a0026215

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. https://doi.org/10.1037/a0020636

JASP Team (2022). JASP (Version 0.16.3) [Computer software].

Jeffreys, H. (1961). *Theory of probability* (3rd ed.). Oxford University Press.

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

Kang, S. H., & Pashler, H. (2012). Learning painting styles: Spacing is advantageous when it promotes discriminative contrast. *Applied Cognitive Psychology, 26*(1), 97–103. https://doi.org/10.1002/acp.1801

Karpicke, J. D., & Aue, W. R. (2015). The testing effect is alive and well with complex materials. *Educational Psychology Review, 27*, 317–326.

Kliegl, O., & Bäuml, K. H. T. (2021). Buildup and release from proactive interference–Cognitive and neural mechanisms. *Neuroscience & Biobehavioral Reviews, 120*, 264–278. https://doi.org/10.1016/j.neubiorev.2020.10.028

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

Kubik, V., Jönsson, F. U., de Jonge, M., & Arshamian, A. (2020). Putting action into testing: Enacted retrieval benefits long-term retention more than covert retrieval. *Quarterly Journal of Experimental Psychology: Human Experimental Psychology, 73*(12), 2093–2105. https://doi.org/10.1177/1747021820945560

Leahy, W., & Sweller, J. (2019). Cognitive load theory, resource depletion and the delayed testing effect. *Educational Psychology Review, 31*, 457–478. https://doi.org/10.1007/s10648-019-09476-2

Lee, H. S., & Ahn, D. (2018). Testing prepares students to learn better: The forward effect of testing in category learning. *Journal of Educational Psychology, 110*(2), 203–217. https://doi.org/10.1037/edu0000211

Lee, H. S., & Ha, H. (2019). Metacognitive judgments of prior material facilitate the learning of new material: The forward effect of metacognitive judgments in inductive learning. *Journal of Educational Psychology, 111*(7), 1189–1201. https://doi.org/10.1037/edu0000339

Lee, M. D., & Wagenmakers, E. J. (2014). *Bayesian cognitive modeling: a practical course*. Cambridge University Press. https://doi.org/10.1017/CBO9781139087759

Markant, D. B. (2019). Effects of biased hypothesis generation on self-directed category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 45*(9), 1552–1568. https://doi.org/10.1037/xlm0000671

Mayer, R. E. (Ed.). (2005). *The Cambridge handbook of multimedia learning*. Cambridge University Press. https://doi.org/10.1017/CBO9780511816819

McCloskey, M. E., & Glucksberg, S. (1978). Natural categories: Well defined or fuzzy sets? *Memory & Cognition, 6*(4), 462–472. https://doi.org/10.3758/BF03197480

McDaniel, M. A., Roediger, H. L., & Mcdermott, K. B. (2007). Generalizing test-enhanced learning from the laboratory to the classroom. *Psychonomic Bulletin & Review, 14*, 200–206. https://doi.org/10.3758/BF03194052

McDermott, K. B. (2021). Practicing retrieval facilitates learning. *Annual Review of Psychology, 72*, 609–633. https://doi.org/10.1146/annurev-psych-010419-051019

Meagher, B. J., McDaniel, M. A., & Nosofsky, R. M. (2022). Effects of feature highlighting and causal explanations on category learning in a natural-science domain. *Journal of Experimental Psychology: Applied, 28*(2), 283–313. https://doi.org/10.1037/xap0000369

Middlebrooks, C. D., Murayama, K., & Castel, A. D. (2017). Test expectancy and memory for important information. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 43*(6), 972–985. https://doi.org/10.1037/xlm0000360

Miyatsu, T., Gouravajhala, R., Nosofsky, R. M., & McDaniel, M. A. (2019). Feature highlighting enhances learning of a complex natural-science category. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 45*(1), 1–16. https://doi.org/10.1037/xlm0000538

Miyatsu, T., & McDaniel, M. A. (2019). Adding the keyword mnemonic to retrieval practice: A potent combination for foreign language vocabulary learning? *Memory & Cognition, 47*(7), 1328–1343. https://doi.org/10.3758/s13421-019-00936-2

Morris, C. D., Bransford, J. D., & Franks, J. J. (1977). Levels of processing versus transfer appropriate processing. *Journal of Verbal Learning and Verbal Behavior, 16*, 519–533. https://doi.org/10.1016/S0022-5371(77)80016-9

Murphy, G. L. (2002). *The big book of concepts*. The MIT Press. https://doi.org/10.7551/mitpress/1602.001.0001

Nosofsky, R. M., & McDaniel, M. A. (2019). Recommendations from cognitive psychology for enhancing the teaching of natural-science categories. *Policy Insights from the Behavioral and Brain Sciences, 6*(1), 21–28. https://doi.org/10.1177/2372732218814861

Nosofsky, R. M., Palmeri, T. J., & McKinley, S. C. (1994). Rule-plus exception model of classification learning. *Psychological Review, 101*, 53–79. https://doi.org/10.1037/0033-295X.101.1.53

Nosofsky, R. M., Sanders, C. A., Meagher, B. J., & Douglas, B. J. (2018). Toward the development of a feature-space representation for a complex natural category domain. *Behavior Research Methods, 50*(2), 530–556. https://doi.org/10.3758/s13428-017-0884-8

O'Day, G. M., & Karpicke, J. D. (2021). Comparing and combining retrieval practice and concept mapping. *Journal of Educational Psychology, 113*(5), 986–997. https://doi.org/10.1037/edu0000486

Pastötter, B., & Bäuml, K. H. T. (2014). Retrieval practice enhances new learning: The forward effect of testing. *Frontiers in Psychology, 5*, 1–5. https://doi.org/10.3389/fpsyg.2014.00286

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*(4), 437–447. https://doi.org/10.1016/j.jml.2009.01.004

Rawson, K. A., & Dunlosky, J. (2011). Optimizing schedules of retrieval practice for durable and efficient learning: How much is enough? *Journal of Experimental Psychology: General, 140*(3), 283–302. https://doi.org/10.1037/a0023956

Rosedahl, L. A., Serota, R., & Ashby, F. G. (2021). When instructions don't help: Knowing the optimal strategy facilitates rule-based but not information-integration category learning. *Journal of Experimental Psychology: Human Perception and Performance, 47*(9), 1226–1236. https://doi.org/10.1037/xhp0000940

Rowland, C. A. (2014). The effect of testing versus restudy on retention: a meta-analytic review of the testing effect. *Psychological Bulletin, 140*(6), 1432–1463. https://doi.org/10.1037/a0037559

Schwieren, J., Barenberg, J., & Dutke, S. (2017). The testing effect in the psychology classroom: a meta-analytic perspective. *Psychology Learning & Teaching, 16*(2), 179–196. https://doi.org/10.1177/1475725717695149

Sotola, L. K., & Crede, M. (2021). Regarding class quizzes: a meta-analytic synthesis of studies on the relationship between frequent low-stakes testing and class performance. *Educational Psychology Review, 33*(2), 407–426. https://doi.org/10.1007/s10648-020-09563-9

Sweller, J. (2010). Cognitive load theory: Recent theoretical advances. In J. L. Plass, R. Moreno, & R. Brünken (Eds.), *Cognitive load theory* (pp. 29–47). Cambridge University Press. https://doi.org/10.1017/CBO9780511844744.004

Sweller, J., van Merriënboer, J. J., & Paas, F. (2019). Cognitive architecture and instructional design: 20 years later. *Educational Psychology Review, 31*(2), 261–292. https://doi.org/10.1007/s10648-019-09465-5

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.

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

van den Broek, G. S. E., van Gog, T., Jansen, E., Pleijsant, M., & Kester, L. (2021). Multimedia effects during retrieval practice: Images that reveal the answer reduce vocabulary learning. *Journal of Educational Psychology, 113*(8), 1587–1608. https://doi.org/10.1037/edu0000499

van Gog, T., & Sweller, J. (2015). Not new, but nearly forgotten: The testing effect decreases or even disappears as the complexity of learning materials increases. *Educational Psychology Review, 27*, 247–264. https://doi.org/10.1007/s10648-015-9310-x

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

Whitehead, P. S., Zamary, A., & Marsh, E. J. (2021). Transfer of category learning to impoverished contexts. *Psychonomic Bulletin & Review, 29*, 1035–1044. https://doi.org/10.3758/s13423-021-02031-7

Yang, C., Chew, S. J., Sun, B., & Shanks, D. R. (2019). The forward effects of testing transfer to different domains of learning. *Journal of Educational Psychology, 111*(5), 809–826. https://doi.org/10.1037/edu0000320

Yang, C., Luo, L., Vadillo, M. A., Yu, R., & Shanks, D. R. (2021). Testing (quizzing) boosts classroom learning: a systematic and meta-analytic review. *Psychological Bulletin, 147*(4), 399–435. https://doi.org/10.1037/bul0000309

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

Yang, C., Potts, R., & Shanks, D. R. (2018). Enhancing learning and retrieval of new information: A review of the forward testing effect. *npj Science of Learning, 3*, 8. https://doi.org/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, and Cognition, 44*, 485–492. https://doi.org/10.1037/xlm0000449

Zeithamova, D., & Maddox, W. T. (2006). Dual-task interference in perceptual category learning. *Memory & Cognition, 34*, 387–398. https://doi.org/10.3758/BF03193416