THE TIMING AND FREQUENCY OF INSTRUCTION TO FOSTER NEW LEARNING: COMPARING COGNITIVE AND MOTIVATIONAL STRATEGIES IN CATEGORY LEARNING

Kyosuke Kakinuma[1,2], Keise Izuma[1,3,4]

[1] School of Economics & Management, Kochi University of Technology, Kochi 780-8515, Japan

[2] Japan Society for the Promotion of Science, Tokyo, 102-0083, Japan

[3] Research Center for Mind, Brain, and Behavior, Kochi University of Technology, Japan

[4] Department of Psychology, University of Southampton, UK

Author Note

Correspondence concerning this article should be addressed to Kyosuke Kakinuma and Keise Izuma, School of Economics & Management, Kochi University of Technology, 2-22 Eikokuji-cho, Kochi-shi, Kochi, 780-8515, Japan (email: kkakinuma10@gmail.com, izuma.keise@kochi-tech.ac.jp)

Acknowledgments

This study was supported by two grants-in-aid from Japan Society for the Promotion of Science: 1) Grant Number 23K12879 (to K.K.) and 2) Grant Number JP19K24680 (to K.I.).

Portions of this work (Experiment 1) have been presented at the 45th Annual Meeting of the Cognitive Science Society (Kakinuma & Izuma, 2023). Specifically, The method and results sections of Experiment 1 was partly based on this preliminary report.

The authors used a generative AI tool (ChatGPT, OpenAI) to assist in drafting some English sentences, and in phrasing and polishing the language. The authors take full responsibility for the content, including its accuracy and integrity.

Abstract

Effective instruction should not only enhance immediate performance but also prepare learners to succeed when learning independently in novel situations (new learning). This study examined which timing and frequency strategies of instruction most effectively foster new learning. We compared six strategies, inspired by theories of motivation (e.g., self-determination theory) and cognition (e.g., testing effect), with two control conditions (no-instruction and full-instruction). Across three experiments (total N = 1,522), participants performed a category learning task. They first received instruction through feature highlighting, whose timing and frequency varied across conditions. Subsequently, their performance was assessed in an independent session where they learned new categories without instruction. An adaptive strategy that tailored instruction to learner performance significantly outperformed the control conditions in initial experiments (Experiments 1 and 2). However, this effect was not observed in a comprehensive experiment (Experiment 3), suggesting that adaptivity itself may not have been the beneficial factor. Instead, an overall examination suggested that the beneficial effects were attributable to two underlying instructional patterns shared by the conditions that outperformed the control conditions: (1) providing instruction in successive sequences, and (2) interspersing tests without instruction by restricting the frequency of instruction. Instructional strategies with these two patterns yielded a significant advantage over the control conditions for independently discovering correct knowledge in novel situations, although they did not reliably improve generalization of that knowledge to other items. This integration of instructional patterns appears to foster new learning by balancing the useful information from instruction with the desirable difficulty of testing.

Keywords: category learning, feature highlighting, forward effect of testing, new learning, test-potentiated learning

The Timing and Frequency of Instruction to Foster New Learning: Comparing Cognitive and Motivational Strategies in Category Learning

Introduction

Instruction from an experienced person to a novice learner is one of the foundations of human society. In an effective instructional environment, learners develop the capacity to explore solutions independently—even in novel situations. For example, imagine a novice learner participating in a field study of wild animals alongside an expert. At first, the learner learns from the expert through instruction that highlights which features are relevant for distinguishing between two animal categories. Years later, the learner might conduct a new field study independently in a different region with unfamiliar animals. The learner can actively explore such uncertain environments and acquire new knowledge by learning from the feedback arising from their own actions. This raises a central question for educational research: Which timing and frequency strategies of instruction most effectively prepare learners to succeed when learning independently in new situations? This is referred to hereafter as new learning—subsequent learning of new material (Chan et al., 2018). The present study develops various instructional strategies inspired by theories from both motivation and cognition and experimentally tests whether each strategy enhances new learning and which strategy is most effective.

A widely used approach to enhance learners' performance across educational settings is to provide instruction at every opportunity. This full-instruction approach has been shown to enhance performance during instruction (Biele et al., 2009; Rosedahl et al., 2021). However, when learners subsequently face unfamiliar situations, this approach may backfire. Overly detailed instruction can reduce opportunities for exploration and hinder learners' ability to adapt

flexibly in new environments. For example, when learners are directly taught how to operate a novel machine, they are less likely to explore its other possible functions or discover alternative ways of using it on their own (Bonawitz et al., 2011). Similarly, learners who are consistently provided with correct answers achieve lower transfer performance than those who are given opportunities to think for themselves (Pan et al., 2018).

Instructional approach in motivation research

In the field of educational psychology, research on motivation has investigated how learners develop independence. According to self-determination theory (Ryan & Deci, 2017), one of the key motivational factors for learners is feeling effective in their interactions with the environment, often referred to as competence (see also White, 1959). Learners' competence can be supported by various instructional practices (Aelterman et al., 2019), such as clearly communicating expectations and goals (i.e., clarifying) and providing adaptive, needs-based help as they make progress (i.e., guiding). Correlational studies have shown that when learners receive such competence-supportive practices, they demonstrate greater behavioral engagement and enjoyment (Skinner & Belmont, 1993; Patall et al., 2024). Learners who receive such support also tend to use more effective self-regulated learning strategies (Aelterman et al., 2019; Sierens et al., 2009). Furthermore, a meta-analysis by Patall et al. (2024) demonstrated that these practices were significantly and positively associated with students' achievement. It also showed that training teachers to implement competence-supportive practices led to significant improvements in students' achievement.

Among competence-supportive practices, adaptive instruction constitutes a promising approach to realizing optimal timing and frequency of instruction. Adaptive instruction deliberately avoids giving full solutions or detailed instructions. Instead, it scaffolds learners

toward success by dynamically adjusting the amount and type of support to match their current level of understanding (Aelterman et al., 2019; van de Pol et al., 2010). To deliver such contingent and tailored support, instructors must engage in diagnostic processes—including formative assessment and real-time monitoring—to accurately assess learners' performance and determine when and how to intervene (Koedinger & Aleven, 2007; van de Pol et al., 2010). As learners demonstrate improved performance, support is gradually faded, thereby transferring responsibility for task completion to them (Aelterman et al., 2019; van de Pol et al., 2010).

The adaptive instruction strategy—adjusting the timing and frequency of instruction according to learners' performance—may foster new learning. Prior studies on motivation suggest that when teachers provide adaptive support, learners are more likely to engage actively, enjoy tasks, and discover effective learning strategies during instruction (Aelterman et al., 2019; Skinner & Belmont, 1993). Building on these findings, learners who receive adaptive instruction are plausibly more likely to apply what they learned during instruction. Consequently, they may be better able to learn independently in novel contexts.

However, previous research has not empirically tested whether adaptive instruction enhances learners' ability to learn independently in novel situations. The adaptive approach has been extensively incorporated across various areas of educational research—not only motivation (Aelterman et al., 2019; Patall et al., 2024), but also scaffolding (van de Pol et al., 2010), memory (Fiechter & Benjamin, 2019), category learning (Pashler et al., 2013), computer-based scaffolding (Belland et al., 2017), intelligent tutoring systems (Koedinger & Aleven, 2007), and adaptive learning technologies (Aleven et al., 2017). Although these studies have examined the association between adaptive approaches and various educational outcomes, their potential to foster new learning remains unclear.

Instructional approaches in cognitive research

Another influential approach has emerged from cognitive and educational psychology: the testing effect (retrieval practice) (Carpenter et al., 2022; Dunlosky et al., 2013). In this line of research, testing is regarded not merely as a tool for assessment but as a powerful learning activity in itself (Karpicke & Roediger, 2008). Traditionally, research on testing has focused on the backward testing effect, whereby retrieving previously studied information enhances retention of that content (Karpicke & Blunt, 2011; Roediger & Karpicke, 2006; Rowland, 2014). More recently, however, researchers have identified another noteworthy benefit, known as the forward testing effect—also referred to as test-potentiated new learning or the interim test effect. This effect shows that taking a test on certain material can enhance the learning of subsequent new material (see reviews by Chan et al., 2018; Pan et al., 2018; Yang et al., 2018). This effect is thought to occur because testing increases learners' engagement in subsequent learning and encourages them to adopt more effective learning strategies (Chan et al., 2018; Yang et al., 2022).

Several studies have demonstrated the forward testing effect using various learning procedures, including word list memory (Pastötter et al., 2011), video lectures (Szpunar et al., 2013), and category learning (Lee & Ahn, 2018). For example, in a category learning study by Lee and Ahn. (2018), participants learned the painting styles of various artists in two separate sessions (A and B). In Session A, as part of the initial learning session, participants were shown a painting with the artist's name one by one (study trials). They were then randomly assigned to a restudy condition or an interim test condition. Participants in the restudy condition were re-presented with the same painting-and-name pairs. In the interim test condition, participants were shown each painting and prompted to type the corresponding artist's name, followed by

corrective feedback. Subsequently, in Session B, as part of the new learning session, all participants studied the painting styles of entirely new artists, again by viewing paintings from each artist. Finally, their performance was assessed with a final test that required them to classify previously unseen paintings from the Session B artists by selecting the correct name from a list of all artists. The results showed that participants in the interim test condition achieved significantly higher accuracy on these final test items than those in the restudy condition. This finding suggests that the interim test in the initial learning (Session A) enhanced the new learning (Session B).

Although testing is an effective learning approach (Chan et al., 2018; Lee & Ahn, 2018; Yang et al., 2018), little is known about its combination with explicit instruction. In educational settings, for example, while students take a practice test, teachers may assist them by highlighting key features relevant to categorization. Such instruction may complement the testing effect by providing useful information to facilitate subsequent learning (Miyatsu et al., 2019). Nevertheless, instruction carries the risk of diminishing the desirable difficulty necessary for learning (Bjork & Bjork, 2011). Indeed, Kang et al. (2023) found that providing instruction on every trial, either during the test or immediately after it as feedback, did not enhance performance compared to testing alone. They suggested that when instruction makes it easier to retrieve answers, learners may have exerted less cognitive effort. This reduced effort may, in turn, prevent them from learning sufficiently from the instruction. These considerations suggest that balancing the benefits of instruction with the desirable difficulty of testing without instruction is critical for new learning. A key open question, therefore, is how the frequency and timing of instruction should be determined to achieve this balance.

The present study

The goal of the present study was to examine which timing and frequency strategy of instruction leads to the best performance when learners subsequently work independently in a novel situation. To this end, we developed and systematically compared several instructional strategies, each grounded in theories of motivation and cognition. To evaluate these strategies, we designed an experimental paradigm consisting of two types of sessions: a teaching session and an independent session. The teaching session implemented the experimental manipulation, in which the timing and frequency of instruction varied across experimental conditions. The independent session assessed the effects of this manipulation on new learning, in which participants worked without instruction.

To examine how learners explore uncertain environments and form new concepts, we employed a category learning task—a type of conceptual learning paradigm (Zeithamova et al., 2019). In this task, participants were presented with a stimulus on each trial and asked to classify it into one of two categories. The task consisted of two phases: a learning phase and a generalization phase. In the learning phase, participants received trial-by-trial feedback (i.e., correct or incorrect). This phase required participants to identify features relevant to the underlying category structure through repeated classification and feedback (Kruschke, 1992; Nosofsky, 1986). In the generalization phase, participants were presented with novel items of the same category structure as in the learning phase. No feedback was provided during this phase. This phase required participants to apply the category rule they had acquired during the learning phase (Bowman & Zeithamova, 2018).

We conducted this category learning task across both teaching and independent sessions. Each session included both the learning and generalization phases, except that in Experiment 1, the teaching sessions did not include the generalization phase. In the teaching session, regardless

of the phase, participants could learn through instruction. Instruction was implemented via feature highlighting, which indicated features relevant to the categories and helped participants identify the underlying category rule (Kang et al., 2023; Miyatsu et al., 2019). The timing and frequency of feature highlighting were experimentally manipulated, such that the pattern of highlighting and no-highlighting trials varied across conditions. In the independent session, participants had to learn new categories solely from the outcomes of their own responses, without any feature highlighting. This self-guided, trial-and-error learning was implemented through trial-by-trial feedback in the learning phase. Based on previous studies on category learning (Bowman et al., 2022; Minda & Smith, 2001), we used classification accuracy in both learning and generalization phases of the independent session as indices of new learning.

Three experiments were conducted. In Experiments 1 and 2 (not preregistered), we tested the effectiveness of an adaptive instruction strategy as a first step. To precisely manipulate this strategy, we developed an algorithm for adaptive instruction. We then compared the adaptive condition with two control conditions to evaluate its effectiveness. One control condition was the no-instruction condition, in which participants did not receive any feature highlighting and thus could not learn from instruction. The other control condition was the full-instruction condition, in which participants received feature highlighting on every trial during the teaching sessions, likely undermining their effort to think on their own. After these experiments, we conducted Experiment 3 (preregistered), in which we examined whether the adaptive strategy offers distinctive benefits over other theoretically promising instructional strategies. In this experiment, we implemented five additional instructional strategies grounded in influential theories of motivation and cognition (e.g., the forward testing effect) and evaluated their effectiveness

alongside that of the adaptive strategy. All experiments were approved by the ethics committee of the first author's institution.

To ensure transparency and reproducibility, all materials, analysis code, data, and codebooks are available on OSF (URL). We report all manipulations and exclusions in this manuscript, and all measures are described either in the manuscript or in the Supplementary Information (SI). In addition, the preregistration of Experiment 3 specified the study design, planned sample size, exclusion criteria, and planned analyses for the primary hypotheses.

Experiments 1 and 2

Experiment 1 served as an initial test of the adaptive strategy. Experiment 2 addressed two limitations identified in Experiment 1 by slightly modifying the procedure and the adaptive algorithm. We report the two experiments together because they shared the same basic structure and outcome measures.

Method

Participants and design. We recruited 172 and 196 adults for Experiments 1 and 2, respectively, via Prolific. After excluding participants who failed attention checks, the final samples consisted of 170 and 195 participants. Participants were randomly assigned to the no-instruction, full-instruction, and adaptive conditions. Detailed participant information is provided in Table 1.

Prior to recruitment, we conducted a power analysis using G*Power (Faul, Erdfelder, Lang, & Buchner, 2007). We set the desired power at .80, the alpha level at .05. As this study was the first to examine the effect of an adaptive strategy using a category learning task, we assumed a medium effect size ($f = .25$) for comparing three between-subjects conditions. The analysis indicated that a sample size of 159 participants would be required.

Materials. We created four types of creature stimuli based on previous category learning studies (Bowman & Zeithamova, 2018; Bozoki et al., 2006; Rosedahl & Ashby, 2018). Each creature type consisted of eight items (see Figure 1). Each item was defined by three features (e.g., crest, foot, and tail), which varied across exemplars within a type (e.g., bird-like creatures). Items were assigned to one of two categories: red or blue. Each category had prototypical features (e.g., blue-category birds had one crest, whereas red-category birds had two). Items were categorized based on the number of features that matched the prototypical features of each category.

Procedure. The experiments were conducted online using jsPsych (de Leeuw, 2015). The experiment consisted of four sessions: three teaching sessions and one independent session (Figure 2). In each session, four types of creatures were randomly assigned, with each creature consisting of eight items. In each trial, an item was presented for up to two seconds, and participants responded by pressing the "f" or "j" key at their own pace. The order of item presentation was pseudo-randomized so that each item appeared once per set, and no more than three items from the same category were shown successively (Bowman & Zeithamova, 2018).

In the teaching sessions of Experiment 1 (Figure 2; upper part), which consisted only of the learning phase, participants classified each of the eight items of a creature (Figure 1) 15 times (a total of 120 trials per session). In each trial, the algorithm determined whether to provide feature highlighting (Figure 2; upper part), which highlighted the relevant features of an item using circles with the corresponding category color (Kang et al., 2023; Miyatsu et al., 2019). The timing and frequency of providing feature highlighting varied depending on the experimental condition. After each classification, they received feedback indicating whether their response was correct or incorrect.

In the independent session, participants classified eight items from a new type of creature, without feature highlighting. This session consisted of two phases: learning and generalization. In the learning phase, participants classified each of the eight items 15 times, receiving feedback after each trial. In the generalization phase, they classified both the same eight items and 40 novel items of the same creature (48 trials in total). The novel items differed slightly in their features from those presented during the learning phase and included additional, irrelevant features. No feedback was provided during this phase. Participants were told to apply the category rule they had learned in the learning phase, but they were not told that irrelevant features were added to the items in the generalization phase.

In Experiment 2, the procedure was identical to that of Experiment 1, except that each teaching session included the generalization phase in addition to the learning phase. Specifically, in the teaching sessions, participants classified each of the eight items 15 times, each followed by feedback. They then proceeded to the generalization phase, in which participants classified both the same eight items and 40 novel items of the same creature, and feedback was not provided for these classifications. During both the learning and generalization phases of the teaching sessions, feature highlighting was provided depending on the condition. We added the generalization phase to allow participants to practice applying category rules to novel stimuli during the teaching sessions.

Experimental conditions. In both Experiments 1 and 2, three experimental conditions were implemented: no-instruction, full-instruction, and adaptive conditions. The conditions differed in the timing and frequency of feature highlighting during the teaching sessions. The independent session was identical across all conditions.

In the no-instruction condition, participants did not receive feature highlighting on any trial. In the full-instruction condition, participants received feature highlighting on every trial during both the learning and generalization phases of the teaching sessions.

In the adaptive condition, the algorithm assessed participants' understanding and adjusted when to provide feature highlighting. Full details are provided in the SI; here we summarize the key steps. To tailor the adaptive teaching process, the algorithm divided the eight items into four pairs based on the structure of the features (Pair 0 to Pair 3 in Figure 1) and performed assessment and adjustment for each pair. At the beginning of the first teaching session, the algorithm provided feature highlighting. It then presented trials without highlighting, requiring participants to respond on their own. Based on participants' responses during these no-highlighting trials, the algorithm assessed their understanding. For the assessment, the algorithm calculated classification accuracy in Experiment 1, whereas Bayesian updating was used in Experiment 2 to enable more timely feature highlighting (see the SI for details). If the assessment met a predetermined performance criterion (accuracy greater than 80% in Experiment 1 and a 95% HDI that did not include the chance level in Experiment 2), the algorithm continued to present a no-highlighting trial. After each no-highlighting trial, the algorithm re-assessed understanding. As long as the re-assessed understanding met the criterion, it withheld feature highlighting and continued the process on a trial-by-trial basis. If the accuracy fell below the criterion, it provided feature highlighting on the next trial. In Experiment 2, this adaptive feature highlighting was applied in the same way in both the learning and generalization phases. Furthermore, in the second and third teaching sessions, the algorithm assessed how many trials were needed for a participant to reach the criterion based on their performance during the preceding session and delayed the start of feature highlighting accordingly.

Measures. Accuracy in the generalization phase was calculated using all 48 trials from the generalization phase of the independent session. Accuracy in the learning phase was calculated using only the latter half of trials from the learning phase (60 trials), because participants were expected to learn the category through trial and error in the earlier part. Additional subsidiary self-reported measures are reported in Table S1.

Data analysis. To examine the effectiveness of the adaptive strategy, we compared it with each of the two control conditions (no-instruction and full-instruction). As the primary analysis, we tested the effects on accuracy in the generalization phase using multiple regression analysis. Two dummy variables were created (Cohen et al., 2003): a no-instruction contrast (adaptive = 0, no-instruction = –1, full-instruction = 0), and a full-instruction contrast (adaptive = 0, no-instruction = 0, full-instruction = –1). In the model, the two dummy variables were entered as independent variables, and accuracy in the generalization phase served as the dependent variable.

As a secondary analysis, we tested the effects on accuracy in the learning phase using a meta-analysis that combined data from both experiments, following the approach recommended by Cumming (2012). Meta-analysis was chosen because both experiments included the same learning phase in the teaching sessions (unlike the generalization phase, which differed between experiments), and pooling the data could increase statistical power. We conducted the meta-analysis using a random-effects model with the R package *metafor* (Viechtbauer, 2010). All statistical tests were one-tailed, because neither the no-instruction nor the full-instruction condition was expected to outperform the adaptive condition. All analyses were performed in R (R Core Team, 2022).

Results and Discussion

Descriptive statistics. The means and 95% confidence intervals (CIs) of accuracy in the learning and generalization phases are shown in Figure 3. Descriptive statistics for other variables and the correlations among variables are provided in Figures S1–S3 and Tables S2–S3 of the SI. The proportion of feature highlighting trials in the adaptive condition is reported in Table S4.

Primary results (accuracy in the generalization phase). In Experiment 1, neither the no-instruction contrast nor the full-instruction contrast was significantly associated with accuracy in the generalization phase ($b^* = .08$, $p = .191$; $b^* = -.02$, $p = .590$; Table S5). We found no evidence supporting the effectiveness of adaptive instruction. In Experiment 2, however, the full-instruction contrast was significantly associated with accuracy ($b^* = .15$, $p = .035$), whereas the no-instruction contrast was not ($b^* = .12$, $p = .070$), as shown in Table S5. This indicates that the adaptive condition exhibited significantly higher accuracy than the full-instruction condition in Experiment 2. These results suggest that by including a generalization phase in the teaching sessions, participants in the performance-adapted condition may have better learned how to apply category rules to novel stimuli.

Secondary results (accuracy in the learning phase). The first meta-analysis compared the adaptive condition with the no-instruction condition and revealed a significant advantage of the adaptive condition ($d = 0.41$, $p = .001$; Figure 4, upper panel). Heterogeneity across studies was low ($\tau^2 = 0.05$, $I^2 = 6.69\%$). The second meta-analysis compared the adaptive condition with the full-instruction condition and again found a significant effect ($d = 0.29$, $p = .013$; Figure 4, lower panel). No heterogeneity was observed ($\tau^2 = 0$, $I^2 = 0\%$). These results indicate that participants in the adaptive condition achieved higher accuracy than both the no-instruction and full-instruction conditions. This suggests that participants in the adaptive condition were better

able to learn categories through trial and error when learning independently in novel situations. We reported the effects of conditions on other variables (Table S6) and the results of additional analyses using an alternative computation of accuracy in the SI (Figure S4 and Table S7).

Additional results regarding accuracy in the generalization phase. In Experiment 2, the adaptive condition exhibited significantly higher accuracy in the generalization phase than the full-instruction condition, but not the no-instruction condition. One possible explanation for the difference in statistical significance is individual variability in ability and engagement with the task. Statistically controlling for baseline variables may reduce standard error and increase the power.

To test this possibility, we simulated baseline accuracy and reaction time. Reaction time was used as an index of task engagement because it was significantly positively correlated with accuracy (Tables S2–S3), suggesting that longer reaction times reflected more careful observation of the stimuli and greater deliberation during categorization. Baseline variables were simulated using data from the first teaching session in the no-instruction condition, which served as a benchmark prior to the experimental manipulation. We then conducted a multiple regression analysis controlling for these variables (see the SI for details). The results showed that, after statistically controlling for these simulated baseline variables, both the no-instruction and full-instruction contrasts were significantly associated with accuracy in the generalization phase (b^* = .12, p = .028; b^* = .15, p = .011; Table S8). These findings suggest that the adaptive condition outperforms the no-instruction condition when baseline variables are statistically controlled.

Experiment 3

Building on the initial evidence for the effectiveness of the adaptive strategy from Experiments 1 and 2, we next examined whether this strategy yields advantages over other

theoretically promising instructional strategies. Thus, in Experiment 3, we introduced five new strategies based on principles from the learning sciences (Table 2), along with the adaptive strategy. By comparing each of these six strategies with the no-instruction and full-instruction conditions, we aimed to comprehensively evaluate which instructional strategy best fosters new learning. In the following paragraphs, we describe the rationale of the five newly introduced strategies.

Among the five new strategies, four of them were based on the forward testing effect: the blocked-50%, blocked-10%, mixed-50%, and mixed-10% conditions (Table 2). Research on the forward testing effect has shown that taking a test on certain material enhances new learning, as noted in the Introduction (Chan et al., 2018). In typical procedures used in prior studies (e.g., Lee & Ahn, 2018), participants were first presented with correct information through study trials. Then, they practiced retrieving it (test condition) or they were re-presented with the same information (restudy condition). Participants in the test condition achieved significantly higher performance in a new learning session than those in the restudy condition (Chan et al., 2018; Lee & Ahn, 2018).

Previous research has proposed several mechanisms for this effect (Chan et al., 2018; Lee & Ahn, 2018; Yang et al., 2019), two of which are particularly relevant to the learning context of the present study. One proposed mechanism is increased task engagement. During test trials, participants may become more aware of the task difficulty and consequently invest greater effort in subsequent learning. Supporting this idea, prior studies have shown that participants in an interim test condition exhibited higher task engagement than those in a condition without interim testing (Healy et al., 2017; Yang et al., 2017). The other mechanism is a shift in learning strategies. While being tested, participants may reevaluate and modify their learning strategies,

switching from less effective to more effective ones. Empirical findings support this idea, showing that participants in an interim test condition used more effective learning strategies than those in a restudy condition (Chan et al., 2018; Yang et al., 2022).

Based on the mechanisms underlying the forward testing effect, the contrast between restudy and test trials in previous research can be mapped onto the contrast between highlighting and no-highlighting trials in the present study. Prior studies have demonstrated the forward testing effect by contrasting test and restudy trials (Chan et al., 2018; Lee & Ahn, 2018; Yang et al., 2019). These studies suggest that test trials make participants aware of the task difficulty, thereby enhancing their task engagement and improving their learning strategies. The contrast between highlighting and no-highlighting trials in the present study is also likely to involve a similar difference in the experience of task difficulty. Although highlighting trials require participants to respond, they are likely easier than no-highlighting trials and may therefore lack desirable difficulty, as is the case for restudy trials. Thus, inserting no-highlighting trials by deliberately limiting the number of highlighting trials may be effective in a manner similar to the forward testing effect. By experiencing the difficulty of retrieval without highlighting, participants may have become more engaged in the task and refined their learning strategies, thereby facilitating performance in subsequent new learning situations. Adapting the standard paradigm used in testing-effect research (Karpicke & Roediger, 2007; Lee & Ahn, 2018), we implemented a blocked-50% strategy (Table 2): trials with feature highlighting and those without it were grouped into separate blocks and alternated in a fixed sequence.

Despite its promise, blocked instruction may pose a risk of metacognitive bias. When learners are presented with a block of highlighted items, they are likely to experience a streak of correct responses, which can make them feel that the task is overly fluent. This sense of fluency

has been shown to inflate learners' confidence (Carpenter et al., 2013), and such overconfidence may, in turn, reduce their task engagement (Dunlosky & Rawson, 2012). Consequently, blocked instruction may undermine engagement during highlighting trials.

To mitigate this potential risk, we introduced three alternative strategies (mixed-50%, blocked-10%, mixed-10%; Table 2). The first was the mixed strategy, which presented highlighting and no-highlighting trials in a mixed sequence, inspired by research on interleaved learning (Kornell & Bjork, 2008; Metcalfe & Xu, 2016). This strategy may reduce overconfidence by limiting the number of successive correct responses. The second was an approach that shortened instructional time and increased tests without instruction (Risko et al., 2024). In our study, this was implemented by providing a small number of highlighting trials and a large number of no-highlighting trials (blocked-10%). This low-frequency instruction strategy may encourage learners to monitor their ability more accurately during opportunities for retrieval practice without highlighting (Scheck & Nelson, 2005; West et al., 2025). Thus, both mixed and low-frequency instruction strategies may help learners maintain high task engagement during highlighting trials while retaining the benefits of the forward testing effect seen in the blocked-50% strategy. Based on these considerations, we included both blocked and mixed strategies and, for each strategy, implemented a low-frequency condition (approximately 10%) as well as a standard-frequency condition (50%). We consider 50% standard because this frequency has often been used in previous research (Karpicke & Roediger, 2007; Lee & Ahn, 2018). Furthermore, by comparing these new conditions to the adaptive condition, we aimed to test whether the effects observed in Experiments 1 and 2 were due to adaptivity itself or merely to low frequency or the mixed presentation of instruction.

In contrast to the strategies in which instructional timing and frequency were externally determined, we also introduced an alternative approach that allowed learners to make their own choices about whether to receive instruction (Choice condition; Table 2). This approach was based on research on motivation and self-regulated learning. According to self-determination theory and research on autonomy support, offering learners choices can benefit motivation (Reeve & Cheon, 2021; Ryan & Deci, 2017). A meta-analysis of the choice effect found that learners who were given choices reported greater task enjoyment and engagement compared to those who were not (Patall et al., 2008). Nevertheless, a potential drawback is that learners may make pedagogically suboptimal choices. Indeed, research on self-regulated learning has shown that learners often prefer less effective strategies (e.g., restudying) over more effective ones (e.g., testing) (Bjork et al., 2013; Rivers, 2021). One way to mitigate this risk is through a stepwise display format, in which questions are presented before additional information (van den Broek et al., 2023). Analogously, in our choice strategy, participants first attempted the task without feature highlighting and then decided whether to receive it.

In summary, we implemented six instructional conditions: blocked-50%, mixed-50%, blocked-10%, mixed-10%, choice, and adaptive. We aimed to test their effects on new learning by comparing each condition with the no-instruction and full-instruction conditions. Furthermore, we planned to exploratorily compare conditions that significantly outperformed both control conditions.

In addition to testing the effectiveness of these six instructional strategies, we examined their potential mediating processes. Specifically, we tested the roles of task engagement, task enjoyment, and learning strategies employed by participants to classify stimuli. Reaction time was used as an index of task engagement, because longer reaction times were associated with

higher accuracy in Experiments 1 and 2 and suggested more careful processing during the task. Task enjoyment was measured with a self-report scale. To estimate category learning strategies, we applied computational models of category learning (Minda & Smith, 2001; Nosofsky, 1987; Shepard, 1957) to participants' trial-by-trial responses.

We preregistered three primary hypotheses. First, the blocked-50%, blocked-10%, mixed-50%, mixed-10%, choice, and adaptive conditions would exhibit higher accuracy in the generalization phase than both the no-instruction and full-instruction conditions. Second, reaction time would mediate the effects of these six strategies.[*1] Third, task enjoyment would mediate the effects of the choice and adaptive strategies.

In addition, we tested secondary hypotheses that were not preregistered. First, the six experimental conditions would exhibit higher accuracy in the learning phase than in both the no-instruction and full-instruction conditions. Second, reaction time and task enjoyment would mediate the effects on accuracy in the learning phase, in the same way as for accuracy in the generalization phase. Third, learning strategies—specifically, prototype use and maximum attention weight, as described in the Data analysis section—would mediate the effects of the six instructional strategies on both accuracy in the learning and generalization phases.

Method

Participants and design. We recruited 1,199 adults via Prolific. According to a preregistered sampling plan, we excluded 42 participants who failed attention check tests or reported on the questionnaire that they took notes or pictures of stimuli many times. The final sample consisted of 1,157 participants. Participants were randomly assigned to one of the eight conditions. Detailed participant information is provided in Table 1.

Prior to recruitment, we estimated the required sample size for Experiment 3. Based on the additional results of Experiment 2, we conducted a power analysis for testing the regression coefficient, assuming that baseline accuracy and reaction time would be statistically controlled (see the SI for details). In the power analysis, the alpha level was set at .05; a one-tailed test was used; and we set the target power for each regression coefficient at .895 so that the joint power (i.e., both the no-instruction and full-instruction contrasts being significant) would be approximately .80 (.895 × .895 = .801). This was because we planned to reject an individual null hypothesis only if both contrasts were significant. The power analysis indicated that 1,096 participants were required.

Materials. To estimate category learning strategy by fitting the models, we modified the stimuli to include six features per item (Figure 5 for an example), instead of three. Using six features made it possible to create items with varying levels of similarity to the categories, enabling an estimation of model fit and parameters. Each creature set consisted of 44 items, with each item characterized by a unique combination of the six features. These were selected from the 64 possible combinations (2^6), excluding 20 items that shared an equal number of features from both categories and thus could not be clearly classified. The structure of the items is shown in Tables S9–S10.

Procedure. The basic procedure was identical to Experiment 2. In Experiment 3, we added a baseline session at the beginning of the experiment to statistically control for baseline categorization accuracy and reaction time. In this session, participants were asked to classify items from a type of creature without feature highlighting, as in the independent session. Due to the limitation of total experiment duration, we reduced the number of teaching sessions from three to two.[*2] Thus, the experiment consisted of four sessions: a baseline session, two teaching

sessions, and an independent session. Each session included both a learning phase and a generalization phase.

Because the number of features increased, we extended the maximum presentation time for each item from two seconds to four seconds. In addition, we adjusted the number of trials in each phase accordingly. In the learning phase of all sessions, participants classified each of 14 items of a creature over six blocks, without repetition within a block (a total of 84 trials). The number of trials in the learning phase was the same across all sessions. In the generalization phase, participants classified 30 novel items of the same creature that were constructed by recombining relevant features from the learning phase. The number of trials in the generalization phase varied across sessions. In the baseline session, the 30 novel items were presented twice (60 trials). In the teaching sessions, the 30 novel items were presented once (30 trials). In the independent session, the 30 novel items and the same 14 items from the learning phase were each presented twice, resulting in 88 trials. The same 14 items were included for the purpose of model fitting.

Experimental conditions. Depending on the condition, the timing and frequency of feature highlighting varied during the teaching sessions. The no-instruction and full-instruction conditions were identical to those used in Experiments 1 and 2.

The adaptive algorithm was nearly identical to that used in Experiments 1 and 2, with minor adjustments reflecting changes in the stimulus structure. The algorithm divided the items into three-item groups based on the number of prototypical features of the creature. Two items with six prototypical features were allocated to Group 1, 12 items with five prototypical features were allocated to Group 2, and 30 items with four prototypical features were allocated to Group 3 (Figure 5). Across all teaching sessions, the algorithm used a Bayesian method to estimate

participants' understanding for each group and determined whether to provide feature highlighting, as in Experiment 2. Furthermore, in the second teaching session, the algorithm calculated the number of trials required for a participant to reach the criterion and delayed the onset of feature highlighting accordingly, as in Experiment 1.

In the blocked-50% and blocked-10% conditions, a block of successive feature highlighting trials and a block of successive no-highlighting trials were presented in an alternating sequence, respectively (Table 3). In the learning phase, which consisted of six 14-trial blocks, the blocked-50% algorithm provided feature highlighting on all trials within the even-numbered blocks (2nd, 4th, and 6th), and withheld feature highlighting in the odd-numbered blocks (1st, 3rd, and 5th). The blocked-10% algorithm provided feature highlighting only during the first four trials of the even-numbered blocks, withholding feature highlighting for the remaining trials. [*3] In the generalization phase, the blocked-50% algorithm alternated between blocks of five feature highlighting trials and five no-highlighting trials, whereas the blocked-10% algorithm provided feature highlighting only on the first trial within each five-trial feature highlighting block.

In the mixed-50% and mixed-10% conditions, feature highlighting trials and no-highlighting trials were presented in a pseudorandom order (Table 3). In the mixed-50% condition, feature highlighting was provided on 50% of all trials, with no more than three successive feature highlighting trials. In the mixed-10% condition, feature highlighting was provided on approximately 10% of all trials and was interspersed with 6 to 10 no-highlighting trials. Participants in the blocked and mixed conditions were told that feature highlighting might be presented during the teaching sessions, but they were not informed about the timing and frequency in which it would appear.

In the choice condition, feature highlighting was provided only upon participants' request. Each trial began with an item presented without feature highlighting for up to four seconds, during which participants could press the L key to request it. If feature highlighting was requested, the item was re-presented with the feature highlighting for the remaining duration. Participants were informed about this option prior to the teaching sessions. This design ensured that the maximum stimulus duration was identical to that in the other conditions, thereby preventing the maximum stimulus duration from becoming a confounding factor.

Measures. Accuracy in the learning and generalization phases was calculated in the same way as in Experiments 1 and 2. Baseline accuracy (i.e., accuracy in the baseline session) was calculated using the same procedure. Reaction time during the teaching sessions was used as an index of task engagement, with mean reaction time calculated after processing outliers (see the SI for details). Baseline reaction time was calculated by averaging reaction times from the baseline session after processing outliers. Task enjoyment was assessed using a four-item subscale of the Intrinsic Motivation Inventory (Ryan, 1982; e.g., "I enjoyed doing the practice-session task very much."; 1 = *not at all*, 7 = *very true*), and a mean score was computed across the four items ($\alpha = .94$). Additional subsidiary self-reported measures are reported in Table S1.

Data analysis. The analyses described below were preregistered, except for the analyses of accuracy in the learning phase, and except for the use of prototype and maximum attention weight as mediators. Any deviations from the preregistered analysis plan are noted in the manuscript. All preregistered analyses are reported either in the main manuscript or in the SI.

To examine the effectiveness of the instructional strategies, we compared six experimental conditions (blocked-50%, blocked-10%, mixed-50%, mixed-10%, choice, and adaptive) against two control conditions (no-instruction and full-instruction), using both

conjunction and individual testing logic (Rubin, 2021). For each comparison, we created two dummy variables: the no-instruction contrast (six experimental conditions = 0, no-instruction condition = –1, full-instruction condition = 0) and the full-instruction contrast (six experimental conditions = 0, no-instruction condition = 0, full-instruction condition = –1). We then conducted separate multiple regression analyses for six subsets of the data, each containing one experimental condition and the two control conditions (e.g., the subset including blocked-50%, no-instruction, and full-instruction). In each regression model, the two dummy variables were entered as independent variables, with baseline accuracy (generalization or learning) and baseline reaction time included as covariates. One-tailed tests were used to assess each regression coefficient. We rejected a null hypothesis about an experimental condition only if the regression coefficients of both dummy variables were significant. We separately tested each experimental condition. Note that although we conducted six multiple regression analyses, we did not need to adjust the alpha level of the tests because we had only one opportunity to make a type 1 error about an individual null hypothesis (Rubin, 2021).

Mediation analyses were performed using the bootstrap method implemented in the R package *mediation* (Tingley et al., 2014), with 10,000 bootstrap samples drawn with replacement and using two-tailed tests.

To estimate the category learning strategies used by participants, we followed procedures established in previous category learning research (Bowman & Zeithamova, 2018). Specifically, we used the prototype model, in which the similarity of each generalization item to the prototype was computed (Bowman & Zeithamova, 2018; Minda & Smith, 2001; Shepard, 1957).[*4] The model was fit to participants' trial-by-trial classification responses in the generalization phase of the independent session. Then, using Monte Carlo simulations, we examined whether the model

fit reliably better than a random model. As a deviation from the preregistration, we did not conduct chi-square tests to compare the proportion of participants who used the prototype strategy across conditions. Instead, we used two estimates derived from the prototype model as mediators indexing category learning strategies (Bowman et al., 2022): (1) prototype use, which indicates the extent to which participants abstracted a central tendency (prototype) for each category and used it to classify items, and (2) maximum attention weight, which reflects the extent to which participants relied heavily on a single feature to classify items. In the present study, the task was designed such that using the prototype and having a lower maximum attention weight (i.e., distributing attention more evenly across features rather than focusing on a single feature) would lead to higher accuracy in the learning and generalization phases. Model fitting and Monte Carlo simulations were performed using MATLAB (MathWorks, Natick, MA).

Results and Discussion

Descriptive statistics. The means and 95% confidence intervals (CIs) of accuracy in the learning and generalization phases are shown in Figure 6. Descriptive statistics for other variables are provided in Figures S5–S10. The correlations among the variables are presented in Table 4. Prototype use was significantly positively correlated with accuracy in the learning and generalization phases ($r = .69$, $p < .001$; $r = .52$, $p < .001$), which indicates that the more participants classified items using the prototype, the higher their accuracy was. In addition, maximum attention weight was significantly negatively correlated with accuracy in the learning and generalization phases ($r = -.46$, $p < .001$; $r = -.46$, $p < .001$), which indicates that the more participants paid attention to all features of items instead of only one of the features, the higher

their accuracy was. The proportion of feature highlighting trials in each condition is reported in Table S11.

Primary results (accuracy in the generalization phase). In the datasets including each of the six experimental conditions, neither the no-instruction contrast nor the full-instruction contrast was significantly associated with accuracy in the generalization phase (*p*s > .053; Table 5). Thus, we found no evidence supporting the prediction that the instructional conditions would yield higher accuracy in the generalization phase than the control conditions.

Secondary results (accuracy in the learning phase). In the dataset including the blocked-10% condition, both the no-instruction and full-instruction contrasts were significantly positively associated with accuracy in the learning phase (b^* = .14, p = .001; b^* = .08, p = .043; Table 6), indicating that the blocked-10% condition yielded significantly higher accuracy than both the no-instruction and full-instruction conditions. In the datasets including the blocked-50% and mixed-50% conditions, respectively, the no-instruction contrast was significantly positively associated with accuracy (b^* = .11, p = .007; b^* = .12, p = .008), whereas the full-instruction contrast was not (b^* = .05, p = .157; b^* = .06, p = .118; Table 6). In the datasets including each of the remaining three conditions, neither the no-instruction contrast nor the full-instruction contrast was significantly associated with accuracy (*p*s > .064; Table 6). In contrast to Experiments 1 and 2, we found no evidence supporting the benefit of the adaptive strategy. Additional analyses using an alternative computation of accuracy are reported in the SI (Table S22).

So far, our data showed that only the blocked-10% strategy enhanced participants' accuracy in the learning phase, compared to both control conditions. To further examine the effects of the blocked-10% strategy on accuracy in the learning phase, we conducted mediation analyses. In the mediation model, four variables were included as mediators: reaction time

during the teaching sessions, task enjoyment, prototype use, and maximum attention weight. Mediation effects through reaction time and maximum attention weight were significant for both the no-instruction and full-instruction contrasts: for reaction time, the standardized indirect effects were .04 (95% CI = [.02, .08], p = .001) and .06 (95% CI = [.02, .10], p = .001), respectively; for maximum attention weight, they were .08 (95% CI = [.05, .12], p < .001) and .03 (95% CI = [.00, .06], p = .048), respectively. In contrast, the mediation effects through task enjoyment and prototype use were not significant (ps > .33). These results suggest that the blocked-10% strategy enhanced participants' task engagement (as indexed by reaction time) and encouraged them to attend to all relevant features of the stimuli, which in turn facilitated their accuracy in the learning phase. The effects of conditions on the mediator variables are reported in Table S12–S15.

Additional results regarding the blocked-10% condition. To investigate whether the effect of the blocked-10% strategy on accuracy in the learning phase was driven by timing, frequency, or both, we compared the blocked-10% condition with each of the mixed-10% and blocked-50% conditions. We conducted a multiple regression analysis on a dataset that included these three conditions. The blocked-10% condition was chosen as the reference condition, and two dummy variables were created: the mixed-10% contrast (blocked-10% = 0, mixed-10% = –1, blocked-50% = 0), and the blocked-50% contrast (blocked-10% = 0, mixed-10% = 0, blocked-50% = –1). In the model, the two dummy variables were entered as independent variables, with baseline accuracy in the learning phase and baseline reaction time included as covariates. Regression coefficients for the two dummy variables were tested using two-tailed tests.

Results showed that the mixed-10% contrast was significantly positively associated with accuracy in the learning phase (b^* = .10, p = .033), whereas the blocked-50% contrast was not (b^*

= .02, p = .719), as shown in Table S16. These results indicate that the blocked-10% condition yielded significantly higher accuracy than the intermixed-10% condition. These findings suggest that timing (i.e., whether feature highlighting is blocked or mixed), rather than frequency, contributed to the effects of the blocked-10% condition on accuracy in the learning phase. This effect was mediated by maximum attention weight (standardized indirect effect = .03, 95% CI [.01, .06], p = .021), but not by any other variable (ps > .14). In addition, results of analyses for the effects on the mediator variables (Tables S17–S18) showed that the blocked-10% condition yielded significantly longer reaction times than the blocked-50% condition (b^* = .07, p = .0496). Specifically, a significant difference was found in reaction times on highlighting trials (b^* = .34, p < .001), but not on no-highlighting trials (b^* = –.05, p = .193). The SI also reports an alternative computation of accuracy (Table S23).

Additional results regarding the adaptive condition. We examined whether adaptive instruction affected reaction time (an index of task engagement), task enjoyment, prototype use, and maximum attention weight (indices of learning strategies), as suggested by previous studies on motivation (Aelterman et al., 2019; Skinner & Belmont, 1993). We selected the mixed-10% condition as an appropriate benchmark because it shares several key characteristics with the adaptive condition—namely, the combination of the feature highlighting and no-highlighting trials, and the no-blocked presentation of feature highlighting. The mixed-10% condition also had the frequency of feature highlighting trials closest to that of the adaptive condition among all conditions (Table S11). In the regression model, a dummy variable (mixed-10% = 0, adaptive = 1) was entered as the independent variable, with baseline accuracy in the generalization phase and baseline reaction time included as covariates. The results showed that the dummy variable was not significantly associated with any of the variables (ps > .30; Tables S19–S21). These

results provide no evidence that adaptive instruction had beneficial effects on task engagement, task enjoyment, or learning strategies.

General Discussion

In the present study, we aimed to examine which timing and frequency strategies of instruction best foster learners' performance when they work independently in novel situations. Across three experiments, we employed a category learning paradigm to systematically evaluate the effectiveness of various instructional strategies.

Effects on accuracy in the learning phase

A meta-analysis of Experiments 1 and 2 showed that the adaptive condition led to significantly higher accuracy in the learning phase than both no-instruction and full-instruction conditions. In contrast, Experiment 3 found no significant effect of the adaptive condition. Instead, the blocked-10% condition yielded significantly higher accuracy than the two control conditions. This effect was mediated by longer reaction time, suggesting increased task engagement, and by lower maximum attention weight, indicating the use of a more effective learning strategy. The blocked-50% and mixed-50% conditions also yielded significantly higher accuracy than the no-instruction condition, with this effect being mediated by longer reaction time and lower maximum attention weight, although the difference between these conditions and the full-instruction condition was not significant.

In Experiment 3, among all conditions, the blocked-10% condition yielded the highest accuracy in the learning phase, followed by the blocked-50% and mixed-50% conditions. These three conditions included the two key instructional patterns: (1) the successive presentation of highlighting trials (which occurred with moderate frequency in the mixed-50% condition) and (2) the alternation between highlighting and no-highlighting trials. In contrast, the instructional

conditions that did not differ significantly from the no-instruction condition contained only one of these instructional patterns. These findings suggest that both instructional patterns may be necessary to produce the effect on new learning. While we had predicted the role of alternation between highlighting and no-highlighting trials based on research on the forward testing effect, we had not anticipated the role of successive highlighting.

The account of the successive presentation of feature highlighting provides a possible explanation for why the results of the adaptive condition differed across the three experiments. In the adaptive condition, the rate of successive feature highlighting differed across the experiments. Specifically, the succession rate of highlighting trials—defined as the number of successive transitions between highlighting trials (e.g., two consecutive trials = 1, three = 2, four = 3) divided by the total number of highlighting trials—was 59% in Experiment 1, 21% in Experiment 2, and 11% in Experiment 3. In other words, feature highlighting in the adaptive condition was most blocked in Experiment 1, intermediate in Experiment 2, and more dispersed in Experiment 3. This raises the possibility that the significant effects observed in the adaptive condition of Experiments 1 and 2 may have resulted not from adaptive instruction per se, but rather from the blocked presentation of feature highlighting. Moreover, in Experiment 3, there were no significant differences between the adaptive and the mixed-10% conditions (the two conditions with similar frequency of feature highlighting) across multiple indices. These findings suggest that the adaptivity of the algorithm, as implemented in our study, may not have provided additional advantages.

The successive presentation of feature highlighting, one of the two key instructional patterns in the present study, can be interpreted in light of research on inductive learning. Based on the psychological mechanism proposed in category learning studies (Carvalho & Goldstone,

2015; 2017), when learners are successively presented with items that share a common attribute, they are likely to attend to the commonalities between these items. According to research on analogical encoding (Gentner et al., 2003), such a focus on commonalities across multiple cases facilitates the acquisition of more abstract and transferable knowledge. In the present study, learners may have been more likely to notice the commonalities that all category-relevant features were highlighted when highlighted trials were presented successively than when they were presented in isolation. This may have led them to abstract the general knowledge that attending to all relevant features was important for accurate classification—a category learning strategy applicable across sessions. As a result, they may have attended to all features in the independent session, as reflected in lower maximum attention weight.

The other key instructional pattern, namely the alternation between highlighting and no-highlighting trials is consistent with findings from the forward testing effect (Chan et al., 2018; Lee & Ahn, 2018; Yang et al., 2019). Prior studies have shown that, as a result of being tested, learners engage more in subsequent learning (Healy et al., 2017; Yang et al., 2017) and switch from less effective to more effective learning strategies (Chan et al., 2018; Yang et al., 2022). In our study, participants may have recognized the difficulty of the task during no-highlighting trials and consequently increased their task engagement, as reflected in longer reaction times. Furthermore, they may have evaluated their prior category learning strategies and acquired more effective category learning strategies during these trials, as indicated by lower maximum attention weight.

To examine the effects of instructional frequency, we compared the blocked-10% and blocked-50% conditions. Although accuracy in the learning phase did not significantly differ between the two conditions, the blocked-10% condition yielded significantly longer reaction

times on highlighting trials than the blocked-50% condition. This suggests that participants in the blocked-10% condition were more engaged with highlighting trials than those in the blocked-50% condition. A possible explanation is that the blocked-10% condition reduced the risk of metacognitive bias that may have been induced in the blocked-50% condition. In the blocked-50% condition, participants may have experienced high fluency from a streak of correct responses during highlighting trials, which can lead to overconfidence (Carpenter et al., 2013) and subsequently reduce task engagement (Dunlosky & Rawson, 2012). In contrast, the blocked-10% condition provided more opportunities for retrieval practice without highlighting, which may have encouraged participants to monitor their ability more accurately. Consequently, participants in the blocked-10% condition may have recognized the difficulty of the task, treated the infrequent feature highlighting as more valuable, and processed it more carefully.

Despite using the same frequency of highlighting as the blocked-10% condition, the intermixed-10% condition did not enhance accuracy in the learning phase. In fact, it yielded significantly lower accuracy in the learning phase than the blocked-10% condition, and this effect was mediated by maximum attention weight. These results suggest that the intermixed-10% condition may have hindered the development of effective learning strategies, leading to reduced accuracy. In previous research using a mixed strategy, exemplars from different categories were intermixed (Kornell & Bjork, 2008; Metcalfe & Xu, 2016), whereas in our study, highlighting and no-highlighting trials were intermixed. This type of intermixing may have made it difficult for learners to compare highlighted items, thereby impairing the formation of an effective learning strategy.

The choice condition showed no advantage over the control conditions. One likely reason is that each trial in this condition initially displayed an item without highlighting. Although we

adopted this type of choice design to enhance accuracy by reducing the risk that participants would overuse feature highlighting, it may have eliminated the benefits of successive highlighting trials. During category learning, learners tend to compare the current item with the previously studied one (Carvalho & Goldstone, 2015; 2017). In our study, participants in the choice condition may have compared a highlighted item with a no-highlighted item rather than with another highlighted item. Consequently, they may have failed to attend to the common principle underlying highlighted items, which may have made it difficult for them to identify an effective category learning strategy.

Effects on accuracy in the generalization phase

Whereas the effects on accuracy in the learning phase were observed across all experiments, the effects on accuracy in the generalization phase differed across experiments. In Experiment 1, no significant differences were found between the adaptive and control conditions. In Experiment 2, the adaptive condition showed significantly higher accuracy than the full-instruction condition in the original regression model, and a significant advantage over the no-instruction condition only when controlling for simulated baseline variables. In Experiment 3, no significant differences across conditions were found.

One possible explanation for the difference in accuracy between the learning and generalization phases is that the generalization phase involved different types of cognitive processing than the learning phase. In the learning phase, learners needed to identify relevant features and determine category assignments based on feedback (Kruschke, 1992; Nosofsky, 1986). In contrast, the generalization phase required them to make judgments based on category representations constructed during the preceding learning phase (Bowman & Zeithamova, 2018).

Feature highlighting, which visually highlighted relevant features, may have been helpful for identifying relevant features, but not for constructing accurate category representations.

Moreover, the differences in accuracy in the generalization phase across experiments may have been due to differences in the generalization items. In Experiments 1 and 2, the items included additional irrelevant features, while retaining the same relevant features as those used in the learning phase. In contrast, the generalization items in Experiment 3 were created by recombining relevant features from the learning phase. Thus, the generalization items in Experiment 3 required participants to rely more heavily on the category representation constructed during the preceding learning phase, potentially limiting the benefits of feature highlighting. Identifying instructional strategies that effectively enhance generalization remains a critical direction for future research.

The theoretical and practical implications

Previous studies on category learning have examined the effectiveness of feature highlighting (Do et al., 2023; Miyatsu et al., 2019; Meagher et al., 2022; Whitehead et al., 2022). Specifically, Kang et al. (2023) investigated its effects on new learning but found no significant benefit. Our findings suggest that the effectiveness of feature highlighting for new learning may depend on how it is presented. The blocked-10%, blocked-50%, and mixed-50% conditions in Experiment 3 as well as the adaptive condition in Experiments 1 and 2 showed significantly higher accuracy in the learning phase compared to the no-instruction condition. These conditions combined the successive presentation of highlighting trials with the alternation between highlighting and no-highlighting trials, which was not incorporated in previous studies. This combination may have helped learners to abstract an effective category learning strategy and refine their existing strategies, thereby enhancing their ability to learn new categories.

Nevertheless, it should be noted that our study used artificial stimuli with well-defined category structures, whereas previous studies employed natural categories with more complex and fuzzy boundaries (e.g., Kang et al., 2023; Miyatsu et al., 2019). Future research should test the effectiveness of this combination with more complex and naturalistic categories.

In addition, our findings have implications for the attentional mechanisms underlying the effects of feature highlighting. A prior study proposed that feature highlighting may facilitate the learning of appropriate attention weights to category-relevant dimensions, thereby enhancing classification accuracy (Miyatsu et al., 2019). However, this possibility has not been directly tested. In the present study, we examined the role of attention weights by fitting a category learning model to participants' responses. Our results showed that maximum attention weight mediated the difference in accuracy in the learning phase between the blocked-10% and no-instruction conditions, whereas prototype use did not. These results provide preliminary evidence for the proposed mechanism.

The present study contributes to research on the testing effect by combining instruction with testing. Recent studies on the testing effect have examined its combination with other techniques (Latimier et al., 2021; McDaniel, 2023), but the combination with other techniques does not necessarily produce stronger effects than testing alone (Kang et al., 2023; O'Day & Karpicke, 2021). For example, Kang et al. (2023) implemented a test-plus-instruction condition, in which instruction (feature highlighting) was provided on every trial either during or immediately after the test, and compared it with a test-only condition. They found no significant difference in performance in the new learning session between them. In our study, the full-instruction condition, despite providing instruction most frequently, did not yield significantly higher accuracy than the no-instruction condition (Table S24). Instead, it led to shorter reaction

times than all other conditions (Table S12), suggesting a decrease in engagement. While instruction provides useful information, it may diminish the desirable difficulty inherent in retrieval effort (Bjork & Bjork, 2011). When learners consistently receive instruction, they may extract only the information needed to answer correctly during instruction, and thus fail to learn from it sufficiently to succeed when working independently in novel situations.

Going beyond previous findings, our study demonstrates that arranging instruction (feature highlighting) and testing without instruction (no feature highlighting) with appropriate timing and frequency plays a crucial role in fostering new learning, as reflected in accuracy in the learning phase. In our study, the blocked-10%, blocked-50%, and mixed-50% conditions improved accuracy in the learning phase compared to the no-instruction condition, and the blocked-10% condition also outperformed the full-instruction condition. These results highlight two critical aspects of the timing and frequency: (1) interspersing tests without instruction by deliberately restricting the frequency of instruction and (2) providing instruction in successive sequences before testing. This appears to balance between the useful information provided by instruction and the desirable difficulty of testing without instruction, enabling learners to acquire knowledge that facilitates subsequent new learning.

Although our study did not employ the exact control conditions used in prior research on the forward testing effect, this difference can be considered a strength rather than a limitation. Specifically, we compared the experimental conditions against an active control condition that involved testing (the no-instruction condition). In contrast, prior studies have typically used a restudy control condition, in which learners were passively provided with information and did not take any tests. Importantly, the active control condition in our study set a higher bar for demonstrating effects because test conditions have been shown to produce better performance in

new learning than restudy conditions (Kang et al., 2023; Lee & Ahn, 2018). Against the strong control condition (the no-instruction condition), the blocked-10% strategy still produced significantly higher accuracy in the learning phase. This suggests that its effect size would likely be even larger if tested against a traditional restudy control condition.

Our findings on the effective combination of instruction and testing may inform instructional practices for real-world educational settings. In school, students learn about the behavior of living creatures in biology classes. Later, they may independently explore the behavior of unfamiliar creatures outside the classroom, using the creatures' responses as feedback. For example, a teacher might present several different examples and successively point out the key characteristics of each one, and then alternate this block of instruction with short quizzes. This combined approach may allow students to extract effective learning strategies from the instruction and subsequently apply them when exploring the behavior of unfamiliar creatures on their own.

The present findings for the adaptive condition suggest that the effects of adaptive instruction reported in previous research on motivation may have been confounded with other pedagogical factors (e.g., instructional methods). Adaptive instruction has been regarded as one of several important practices in motivation research (Aelterman et al., 2019). The existing evidence for its effect has come from surveys or interventions that combined multiple instructional techniques, as reflected in a recent meta-analysis (Patall et al., 2023). However, these methodologies were limited in their internal validity. Our findings point to a potential confound that the positive effects of the adaptive condition in Experiments 1 and 2 might be attributable to its incidental use of a more blocked presentation compared to Experiment 3. This suggests that, more broadly, positive outcomes often attributed to adaptive instruction in less

controlled settings may in fact be driven by confounding strategies—such as blocked presentation—rather than by adaptivity itself. Future research should use experimental methods to isolate the unique effect of adaptive instruction from other confounding pedagogical factors.

In the present study, the adaptive instruction algorithm has several aspects that could be improved. First, the algorithm relied on only two indicators to estimate learners' understanding: classification accuracy and the number of trials needed to reach a mastery level. However, in real world settings, a wider range of methods can be used to assess learner understanding. For example, teachers may infer understanding based on facial expressions, tone of voice, and reaction time. Incorporating a broader range of estimation methods may enable more precise assessments and allow the algorithm to support learners more appropriately in timing and frequency.

Second, the algorithm adjusted only partial aspects of instruction. Specifically, the adaptive algorithm adjusted for the timing of providing feature highlighting and for which items feature highlighting was presented. However, it did not adjust the quality of instruction. For example, it neither varied the number of highlighted features, nor employed different types of instruction (e.g., explanations of category rules or prompts addressing common errors). Such diverse and flexible adjustments may make adaptive instruction more attuned to learners' needs.

Third, the simplicity of the experimental task constrained the reliability and breadth of the indicators used for the algorithm. In the experimental task, we employed a binary-choice format in which participants could respond correctly by chance even without a full understanding of the category. In contrast, response formats with more options or open-ended responses may offer a more accurate assessment of learners' understanding. Moreover, real-world learning situations often involve more complex and heterogeneous content than our experimental task.

Under such conditions, the algorithm may become more effective if it incorporates more granular, multi-step adjustments to accommodate the substantial variability across individual learners. Refining these aspects in future research may help improve the design of adaptive strategies and yield more definitive implications for motivation research.

A limitation of this study is that we examined our instructional strategies only within a category learning paradigm, leaving it unclear whether the findings can be extended to other learning contexts. Future research should therefore investigate the effectiveness of a blocked-10% strategy in promoting new learning in other tasks. For example, using a probabilistic reversal learning task, we could test its effect on learners' ability to infer the abstract rule in new situations (Hampton et al., 2006; Marković et al., 2019). In this task, learners choose between two options with different reward probabilities that periodically reverse. This requires learners to identify the underlying rule by which the reward probabilities switch. A blocked-10% strategy could provide a block of infrequent trials highlighting the currently optimal choice, followed by a block of no-highlighting trials. Compared to no-instruction or full-instruction, this strategy may enhance both task engagement and the acquisition of an effective learning strategy for identifying the underlying rule, which, in turn, may foster better performance when learners encounter a new underlying rule.

Conclusion

In the present study, we investigated which timing and frequency strategies best foster new learning. Our findings demonstrated that the blocked-10% strategy—which involved the successive presentation of instructions, an increased frequency of tests without instruction, and alternation between them—fostered the ability to independently discover correct knowledge in novel situations, but did not enhance generalization. This effect is likely driven by increased task

engagement and the development of effective learning strategies during instruction. By contrast, the adaptive strategy—which dynamically adjusts instruction according to learners' performance—did not yield significant effects under the conditions of the present study. To foster new learning, teachers might consider alternating between providing successive instructions and giving learners more opportunities to work on their own.

References

Aelterman, N., Vansteenkiste, M., Haerens, L., Soenens, B., Fontaine, J. R. J., & Reeve, J. (2019). Toward an integrative and fine-grained insight in motivating and demotivating teaching styles: The merits of a circumplex approach. *Journal of Educational Psychology, 111*(3), 497–521. https://doi.org/10.1037/edu0000293

Aleven, V., McLaughlin, E. A., Glenn, R. A., & Koedinger, K. R. (2017). Instruction based on adaptive learning technologies. In R. E. Mayer & P. Alexander (Eds.), *Handbook of research on learning and instruction* (2nd ed., pp. 522–560). Routledge.

Belland, B. R., Walker, A. E., Kim, N. J., & Lefler, M. (2017). A Bayesian network meta-analysis to synthesize the influence of contexts of scaffolding use on cognitive outcomes in STEM education. *Review of Educational Research, 87*(6), 1042–1081. https://doi.org/10.3102/0034654317723009

Biele, G., Rieskamp, J., & Gonzalez, R. (2009). Computational models for the combination of advice and individual learning. *Cognitive Science, 33*(2), 206–242. https://doi.org/10.1111/j.1551-6709.2009.01010.x

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* (2nd ed., pp. 59–68). Worth Publishers.

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

Bonawitz, E., Shafto, P., Gweon, H., Goodman, N. D., Spelke, E., & Schulz, L. (2011). The double-edged sword of pedagogy: Instruction limits spontaneous exploration and discovery. *Cognition, 120*(3), 322–330. https://doi.org/10.1016/j.cognition.2010.10.001

Bowman, C. R., Iwashita, T., & Zeithamova, D. (2022). The effects of age on category learning and prototype- and exemplar-based generalization. *Psychology and Aging, 37*(7), 800–815. https://doi.org/10.1037/pag0000714

Bowman, C. R., & Zeithamova, D. (2018). Abstract memory representations in the ventromedial prefrontal cortex and hippocampus support concept generalization. *Journal of Neuroscience*, *38*(10), 2605-2614. https://doi.org/10.1523/JNEUROSCI.2811-17.2018

Bowman, C. R., & Zeithamova, D. (2020). Training set coherence and set size effects on concept generalization and recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 46*(8), 1442–1464. https://doi.org/10.1037/xlm0000824

Bozoki, A. C., Grossman, M., & Smith, E. E. (2006). Can patients with Alzheimer's disease learn a category implicitly? *Neuropsychologia*, *44*(5), 816-827. https://doi.org/10.1016/j.neuropsychologia.2005.08.001

Carpenter, S. K., Pan, S. C., & Butler, A. C. (2022). The science of effective learning with spacing and retrieval practice. *Nature Reviews Psychology*, *1*(9), 496–511. https://doi.org/10.1038/s44159-022-00089-1

Carpenter, S. K., Wilford, M. M., Kornell, N., & Mullaney, K. M. (2013). Appearances can be deceiving: Instructor fluency increases perceptions of learning without increasing actual learning. *Psychonomic Bulletin & Review*, *20*(6), 1350–1356. https://doi.org/10.3758/s13423-013-0442-z

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

Carvalho, P. F., & Goldstone, R. L. (2017). The sequence of study changes what information is attended to, encoded, and remembered during category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *43*(11), 1699–1719. https://doi.org/10.1037/xlm0000406

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

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

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences* (3rd ed.). Lawrence Erlbaum Associates.

Cumming, G. (2012). *Understanding the new statistics: Effect sizes, confidence intervals, and meta-analysis*. Routledge.

de Leeuw, J. R. (2015). jsPsych: A JavaScript library for creating behavioral experiments in a Web browser. *Behavior Research Methods, 47*(1), 1–12. https://doi.org/10.3758/s13428-014-0458-y

Do, L. A., & Thomas, A. K. (2023). The underappreciated benefits of interleaving for category learning. *Journal of Intelligence*, *11*(8), 153. https://doi.org/10.3390/jintelligence11080153

Dunlosky, J., & Rawson, K. A. (2012). Overconfidence produces underachievement: Inaccurate self evaluations undermine students' learning and retention. *Learning and Instruction*, *22*(4), 271–280. https://doi.org/10.1016/j.learninstruc.2011.08.003

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

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

Fiechter, J. L., & Benjamin, A. S. (2019). Techniques for scaffolding retrieval practice: The costs and benefits of adaptive versus diminishing cues. *Psychonomic Bulletin & Review, 26*(5), 1666–1674. https://doi.org/10.3758/s13423-019-01617-6

Gentner, D., Loewenstein, J., & Thompson, L. (2003). Learning and transfer: A general role for analogical encoding. *Journal of Educational Psychology, 95*(2), 393–408. https://doi.org/10.1037/0022-0663.95.2.393

Hampton, A. N., Bossaerts, P., & O'doherty, J. P. (2006). The role of the ventromedial prefrontal cortex in abstract state-based inference during decision making in humans. *Journal of Neuroscience*, *26*(32), 8360-8367. https://doi.org/10.1523/JNEUROSCI.1010-06.2006

Healy, A. F., Jones, M., Lalchandani, L. A., & Tack, L. A. (2017). Timing of quizzes during learning: Effects on motivation and retention. *Journal of Experimental Psychology: Applied*, *23*(2), 128. https://doi.org/10.1037/xap0000123

Kang, Y., Ha, H., & Lee, H. S. (2023). When more is not better: Effects of interim testing and feature highlighting in natural category learning. *Educational Psychology Review*, *35*(2), 51. https://doi.org/10.1007/s10648-023-09772-y

Karpicke, J. D., & Blunt, J. R. (2011). Retrieval practice produces more learning than elaborative studying with concept mapping. *Science*, *331*(6018), 772–775. https://doi.org/10.1126/science.1199327

Karpicke, J. D., & Roediger, H. L., III. (2007). Repeated retrieval during learning is the key to long-term retention. *Journal of Memory and Language*, *57*(2), 151–162. https://doi.org/10.1016/j.jml.2006.09.004

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

Koedinger, K. R., & Aleven, V. (2007). Exploring the assistance dilemma in experiments with cognitive tutors. *Educational Psychology Review, 19*(3), 239–264. https://doi.org/10.1007/s10648-007-9049-0

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

Kruschke, J. K. (1992). ALCOVE: A connectionist model of human category learning. *Psychological Review, 99*(1), 22–44. https://doi.org/10.1037/0033-295X.99.1.22

Latimier, A., Peyre, H., & Ramus, F. (2021). A meta-analytic review of the benefit of spacing out retrieval practice episodes on retention. *Educational Psychology Review*, *33*(3), 959–987. https://doi.org/10.1007/s10648-020-09572-8

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

Marković, D., Reiter, A. M. F., & Kiebel, S. J. (2019). Predicting change: Approximate inference under explicit representation of temporal structure in changing environments. *PLOS Computational Biology, 15*(1), e1006707. https://doi.org/10.1371/journal.pcbi.1006707

McDaniel, M. A. (2023). Combining retrieval practice with elaborative encoding: Complementary or redundant? *Educational Psychology Review, 35*(3), Article 75. https://doi.org/10.1007/s10648-023-09784-8

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

Metcalfe, J., & Xu, J. (2016). People mind wander more during massed than spaced inductive learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 42*(6), 978–984. https://doi.org/10.1037/xlm0000216

Minda, J. P., & Smith, J. D. (2001). Prototypes in category learning: The effects of category size, category structure, and stimulus complexity. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 27*(3), 775–799. https://doi.org/10.1037/0278-7393.27.3.775

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

Nosofsky, R. M. (1986). Attention, similarity, and the identification–categorization relationship. *Journal of Experimental Psychology: General, 115*(1), 39–57. https://doi.org/10.1037/0096-3445.115.1.39

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

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

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. https://doi.org/10.1037/a0031679

Pastötter, B., & Bäuml, K.-H. T. (2014). Retrieval practice enhances new learning: The forward effect of testing. *Frontiers in Psychology, 5*, 83305. https://doi.org/10.3389/fpsyg.2014.00286

Patall, E. A., Cooper, H., & Robinson, J. C. (2008). The effects of choice on intrinsic motivation and related outcomes: A meta-analysis of research findings. *Psychological Bulletin, 134*(2), 270–300. https://doi.org/10.1037/0033-2909.134.2.270

Patall, E. A., Yates, N., Lee, J., Chen, M., Bhat, B. H., Lee, K., Beretvas, S. N., Lin, S., Yang, S. M., Jacobson, N. G., Harris, E., & Hanson, D. J. (2024). A meta-analysis of teachers' provision of structure in the classroom and students' academic competence beliefs, engagement, and achievement. *Educational Psychologist, 59*(1), 1–26. https://doi.org/10.1080/00461520.2023.2274104

R Core Team. (2022). *R: A language and environment for statistical computing* (Version 4.2.0) [Computer software]. R Foundation for Statistical Computing. https://www.R-project.org/

Reeve, J., & Cheon, S. H. (2021). Autonomy-supportive teaching: Its malleability, benefits, and potential to improve educational practice. *Educational Psychologist, 56*(1), 54–77. https://doi.org/10.1080/00461520.2020.1862657

Risko, E. F., Liu, J., & Bianchi, L. (2024). Speeding lectures to make time for retrieval practice: Can we improve the efficiency of interpolated testing? *Journal of Experimental Psychology: Applied, 30*(2), 268–281. https://doi.org/10.1037/xap0000494

Rivers, M. L. (2021). Metacognition about practice testing: A review of learners' beliefs, monitoring, and control of test-enhanced learning. *Educational Psychology Review, 33*(3), 823–862. https://doi.org/10.1007/s10648-020-09578-2

Roediger, H. L., III, & Karpicke, J. D. (2006). The power of testing memory: Basic research and implications for educational practice. *Perspectives on Psychological Science, 1*(3), 181–210. https://doi.org/10.1111/j.1745-6916.2006.00012.x

Rosedahl, L., & Ashby, F. G. (2018). *A new stimulus set for cognitive research*. PsyArXiv. https://doi.org/10.17605/OSF.IO/2XZ3Q

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

Rubin, M. (2021). When to adjust alpha during multiple testing: A consideration of disjunction, conjunction, and individual testing. *Synthese, 199*(3), 10989–11018. https://doi.org/10.1007/s11229-021-03276-4

Ryan, R. M. (1982). Control and information in the intrapersonal sphere: An extension of cognitive evaluation theory. *Journal of Personality and Social Psychology, 43*(3), 450–461. https://doi.org/10.1037/0022-3514.43.3.450

Ryan, R. M., & Deci, E. L. (2017). *Self-determination theory: Basic psychological needs in motivation, development, and wellness.* The Guilford Press.

Scheck, P., & Nelson, T. O. (2005). Lack of pervasiveness of the underconfidence-with-practice effect: Boundary conditions and an explanation via anchoring. *Journal of Experimental Psychology: General, 134*(1), 124–128. https://doi.org/10.1037/0096-3445.134.1.124

Shepard, R. N. (1957). Stimulus and response generalization: A stochastic model relating generalization to distance in psychological space. *Psychometrika, 22*(4), 325–345. https://doi.org/10.1007/BF02288967

Sierens, E., Vansteenkiste, M., Goossens, L., Soenens, B., & Dochy, F. (2009). The synergistic relationship of perceived autonomy support and structure in the prediction of self-regulated learning. *The British Journal of Educational Psychology, 79*(1), 57–68. https://doi.org/10.1348/000709908X304398

Skinner, E. A., & Belmont, M. J. (1993). Motivation in the classroom: Reciprocal effects of teacher behavior and student engagement across the school year. *Journal of Educational Psychology, 85*(4), 571–581. https://doi.org/10.1037/0022-0663.85.4.571

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

Tingley, D., Yamamoto, T., Hirose, K., Keele, L., & Imai, K. (2014). mediation: R package for causal mediation analysis. *Journal of Statistical Software, 59*(5), 1–38. https://doi.org/10.18637/jss.v059.i05

van den Broek, G. S. E., Gerritsen, S. L., Oomen, I. T. J., Velthoven, E., van Boxtel, F. H. J., Kester, L., & van Gog, T. (2023). Optimizing multiple-choice questions for retrieval practice: Delayed display of answer alternatives enhances vocabulary learning. *Journal of Educational Psychology, 115*(8), 1087–1109. https://doi.org/10.1037/edu0000810

van de Pol, J., Volman, M., & Beishuizen, J. (2010). Scaffolding in teacher–student interaction: A decade of research. *Educational Psychology Review, 22*(3), 271–296. https://doi.org/10.1007/s10648-010-9127-6

Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software, 36*(3), 1–48. https://doi.org/10.18637/jss.v036.i03

West, J. T., Kuhns, J. M., Touron, D. R., & Mulligan, N. W. (2025). Increased metamemory accuracy with practice does not require practice with metamemory. *Quarterly Journal of Experimental Psychology, 78*(7), 1280-1302. https://doi.org/10.1177/17470218241269322

White, R. W. (1959). Motivation reconsidered: The concept of competence. *Psychological Review, 66*(5), 297–333. https://doi.org/10.1037/h0040934

Whitehead, P. S., De-Jesús-Echevarría, M., & Egner, T. (2022). Transfer of category learning to impoverished contexts. *Psychonomic Bulletin & Review, 29*(3), 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), 751–763. https://doi.org/10.1037/edu0000321

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*, Article 8. https://doi.org/10.1038/s41539-018-0024-y

Yang, C., Yue, C., Dang, J., Chen, Y., & Shanks, D. R. (2022). Testing potential mechanisms underlying test-potentiated new learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 48*(8), 1127–1143. https://doi.org/10.1037/xlm0001021

Zeithamova, D., Mack, M. L., Braunlich, K., Davis, T., Seger, C. A., Van Kesteren, M. T., & Wutz, A. (2019). Brain mechanisms of concept learning. *Journal of Neuroscience, 39*(42), 8259-8266. https://doi.org/10.1523/JNEUROSCI.1166-19.2019

Footnote

[1] In our preregistration, we initially identified recognition accuracy as a potential mediator. However, we treated recognition accuracy as a dependent variable, consistent with a previous category learning study (Bowman & Zeithamova, 2020). The results of the recognition accuracy analysis are reported in the SI (Table S25), because they fall outside the primary scope of the present study.

[2] Before Experiment 3, we analyzed the data from Experiment 2 and confirmed that accuracy in the no-highlighting trials of the adaptive condition improved from teaching session 1 to session 2, but not from session 2 to session 3 (Figure S11). These results suggest that two sessions were sufficient to capture the effects of the adaptive strategy.

[3] In both the blocked-10% and mixed-10% conditions, the proportion of feature highlighting trials in the learning phase was approximately 14%, because the number of trials in that phase was not a multiple of 10.

[4] We also fit the exemplar model, in which similarity was computed between each generalization item and all exemplars presented during the learning phase (Nosofsky, 1987). However, because a greater number of participants were better fit by the prototype model (Table S26), we used the estimates derived from the prototype model as indices of learning strategies in the present study.

Table 1

Demographic Characteristics and Condition Assignments of Participants

	Experiment 1	Experiment 2	Experiment 3
Recruited (n)	172	196	1,199
Final sample (n)	170	195	1,157
Age (M ± SD)	35.42 ± 7.99	32.88 ± 7.77	34.10 ± 7.75
Female	75	69	607
Male	94	124	542
Other	0	1	6
Not reported	1	1	2
Compensation	£3.80	£4.50	£3.75
Condition assignment (n)			
No-instruction	66	61	142
Full-instruction	53	69	145
Adaptive	51	65	142
Blocked-50%	-	-	146
Blocked-10%	-	-	147
Mixed-50%	-	-	143
Mixed-10%	-	-	151
Choice	-	-	141

Table 2

Overview of the Eight Conditions in Experiment 3

	Condition	Description
1	No-instruction	No feature highlighting was provided on any trial.
2	Full-instruction	Feature highlighting was provided on every trial.
3	Blocked-50%	Feature highlighting and no-feature highlighting trials were grouped into separate blocks and presented in a fixed sequence (50% of trials included feature highlighting).
4	Blocked-10%	Feature highlighting and no-feature highlighting trials were grouped into separate blocks and presented in a fixed sequence (10% of trials included feature highlighting).
5	Mixed-50%	Feature highlighting and no-feature highlighting trials were mixed in a pseudorandom sequence (50% of trials included feature highlighting).
6	Mixed-10%	Feature highlighting and no-feature highlighting trials were mixed in a pseudorandom sequence (10% of trials included feature highlighting).
7	Choice	Participants first attempted each trial without feature highlighting and then decided whether to receive it.
8	Adaptive	The algorithm estimated learners' ability in real time and dynamically adjusted whether to provide feature highlighting.

Table 3

Timing and Frequency of Feature Highlighting in the Blocked and Mixed Conditions

	Learning phase (14 trials per block)		
	Block 1	Block 2	Block 3
Blocked-50%	0,0,0,0,0,0,0,0,0,0,0,0,0,0	1,1,1,1,1,1,1,1,1,1,1,1,1,1	0,0,0,0,0,0,0,0,0,0,0,0,0,0
Blocked-10%	0,0,0,0,0,0,0,0,0,0,0,0,0,0	1,1,1,1,0,0,0,0,0,0,0,0,0,0	0,0,0,0,0,0,0,0,0,0,0,0,0,0
Mixed-50%	0,1,0,0,1,1,0,1,1,1,0,0,0,1	1,0,1,0,0,0,1,1,0,0,0,1,1,1	0,1,1,0,0,1,0,0,1,0,1,1,0,1
Mixed-10%	0,0,0,0,0,0,1,0,0,0,0,1,0,0	0,0,0,0,0,0,1,0,0,0,0,0,0,1	0,0,0,0,0,1,0,0,0,0,0,0,1,0
	Block 4	Block 5	Block 6
Blocked-50%	1,1,1,1,1,1,1,1,1,1,1,1,1,1	0,0,0,0,0,0,0,0,0,0,0,0,0,0	1,1,1,1,1,1,1,1,1,1,1,1,1,1
Blocked-10%	1,1,1,1,0,0,0,0,0,0,0,0,0,0	0,0,0,0,0,0,0,0,0,0,0,0,0,0	1,1,1,1,0,0,0,0,0,0,0,0,0,0
Mixed-50%	0,1,1,1,0,1,1,1,0,0,0,1,0,0	0,1,0,0,1,0,1,1,1,0,0,1,1,0	1,0,0,1,1,0,1,0,0,1,0,1,0,1
Mixed-10%	0,0,0,1,0,0,0,0,1,0,0,0,0,0	1,0,0,0,0,0,0,0,0,1,0,0,0,0	0,0,0,1,0,0,0,0,0,0,1,0,0,0
	Generalization phase (30 trials)		
Blocked-50%	0,0,0,0,0, 1,1,1,1,1, 0,0,0,0,0, 1,1,1,1,1, 0,0,0,0,0, 1,1,1,1,1		
Blocked-10%	0,0,0,0,0, 1,0,0,0,0, 0,0,0,0,0, 1,0,0,0,0, 0,0,0,0,0, 1,0,0,0,0		
Mixed-50%	0,0,1,0,0, 1,1,0,1,1, 1,0,1,0,1, 0,0,1,1,0, 1,1,0,0,1, 1,0,1,0,0		
Mixed-10%	0,0,0,0,0, 0,0,1,0,0, 0,0,0,0,0, 0,0,0,1,0, 0,0,0,0,0, 0,1,0,0,0		

Note. 0 = no-feature highlighting trial; 1 = feature highlighting trial. The timing of feature highlighting was fixed in the blocked conditions and pseudo-randomized in the mixed conditions.

Table 4

Correlations among Variables in Experiment 3

		1	2	3	4	5	6	7	8	9
1	Accuracy in the generalization phase	-								
2	Accuracy in the learning phase	.78	-							
3	Reaction time during teaching sessions	.56	.56	-						
4	Task enjoyment	.21	.17	.21	-					
5	Prototype use	.69	.52	.33	.16	-				
6	Maximum of attention weight	-.46	-.46	-.45	-.15	-.06	-			
7	Baseline accuracy (generalization phase)	.40	.38	.32	.09	.29	-.24	-		
8	Baseline accuracy (learning phase)	.40	.40	.34	.13	.29	-.22	.58	-	
9	Baseline reaction time	.51	.48	.72	.18	.33	-.36	.40	.40	-

Note. All coefficients are significant ($p < .05$).

Table 5

Effects of Each Contrast on Accuracy in the Generalization Phase (Experiment 3)

Experimental Condition	Contrast	b^*	95% CI	*SE*	*t*	*p*
Blocked-50%	No-instruction contrast	.03	[-.06, .12]	.05	0.60	.276
	Full-instruction contrast	-.03	[-.12, .06]	.05	-0.62	.734
Blocked-10%	No-instruction contrast	.05	[-.04, .14]	.05	1.12	.131
	Full-instruction contrast	.00	[-.09, .09]	.05	-0.06	.523
Mixed-50%	No-instruction contrast	.08	[-.02, .17]	.05	1.62	.053
	Full-instruction contrast	.03	[-.07, .12]	.05	0.53	.299
Mixed-10%	No-instruction contrast	-.03	[-.13, .06]	.05	-0.74	.770
	Full-instruction contrast	-.09	[-.18, .00]	.05	-1.89	.970
Choice	No-instruction contrast	-.01	[-.11, .08]	.05	-0.29	.613
	Full-instruction contrast	-.07	[-.17, .02]	.05	-1.58	.942
Adaptive	No-instruction contrast	-.01	[-.10, .08]	.05	-0.23	.591
	Full-instruction contrast	-.07	[-.16, .03]	.05	-1.40	.919

Note. b^* = standardized partial regression coefficient. Each regression was conducted using a data subset that included the listed experimental condition and the two control conditions (no-instruction and full-instruction). Coefficients were tested using one-tailed tests. Control variables (baseline accuracy in the generalization phase and baseline reaction time) were included in the model but are not reported in the table for brevity.

Table 6

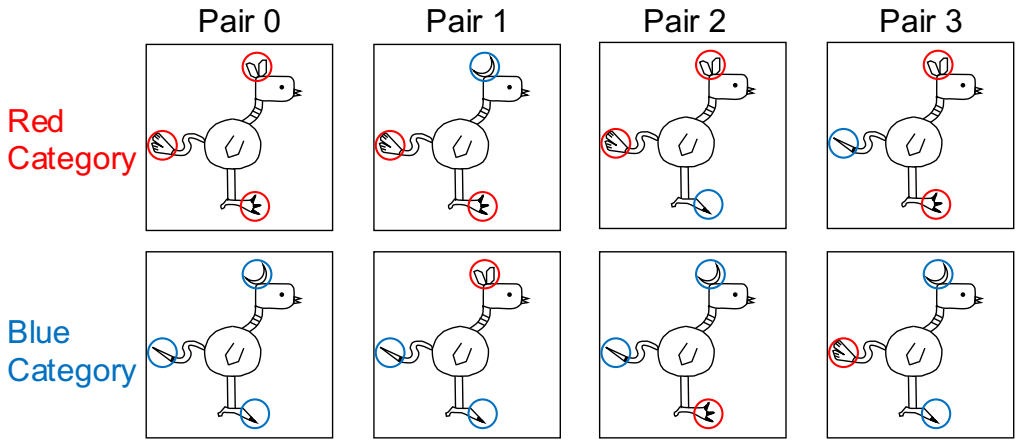
Effects of Each Contrast on Accuracy in the Learning Phase (Experiment 3)

Experimental Condition	Contrast	b^*	95% CI	*SE*	*t*	*p*
Blocked-50%	No-instruction contrast	**.11**	[.02, .20]	.05	2.45	.007
	Full-instruction contrast	.05	[-.04, .14]	.05	1.01	.157
Blocked-10%	No-instruction contrast	**.14**	[.05, .23]	.05	3.11	.001
	Full-instruction contrast	**.08**	[-.01, .17]	.05	1.72	.043
Mixed-50%	No-instruction contrast	**.12**	[.02, .21]	.05	2.44	.008
	Full-instruction contrast	.06	[-.04, .15]	.05	1.19	.118
Mixed-10%	No-instruction contrast	.03	[-.06, .12]	.05	0.63	.264
	Full-instruction contrast	-.04	[-.13, .06]	.05	-0.75	.773
Choice	No-instruction contrast	.07	[-.02, .17]	.05	1.53	.064
	Full-instruction contrast	.01	[-.09, .10]	.05	0.13	.446
Adaptive	No-instruction contrast	.05	[-.04, .15]	.05	1.10	.137
	Full-instruction contrast	-.01	[-.11, .08]	.05	-0.28	.610

Note. b^* = standardized partial regression coefficient. Each regression was conducted using a data subset that included the listed experimental condition and the two control conditions (no-instruction and full-instruction). Coefficients were tested using one-tailed tests. Control variables (baseline accuracy in the learning phase and baseline reaction time) were included in the model but are not reported in the table for brevity.

Figure 1

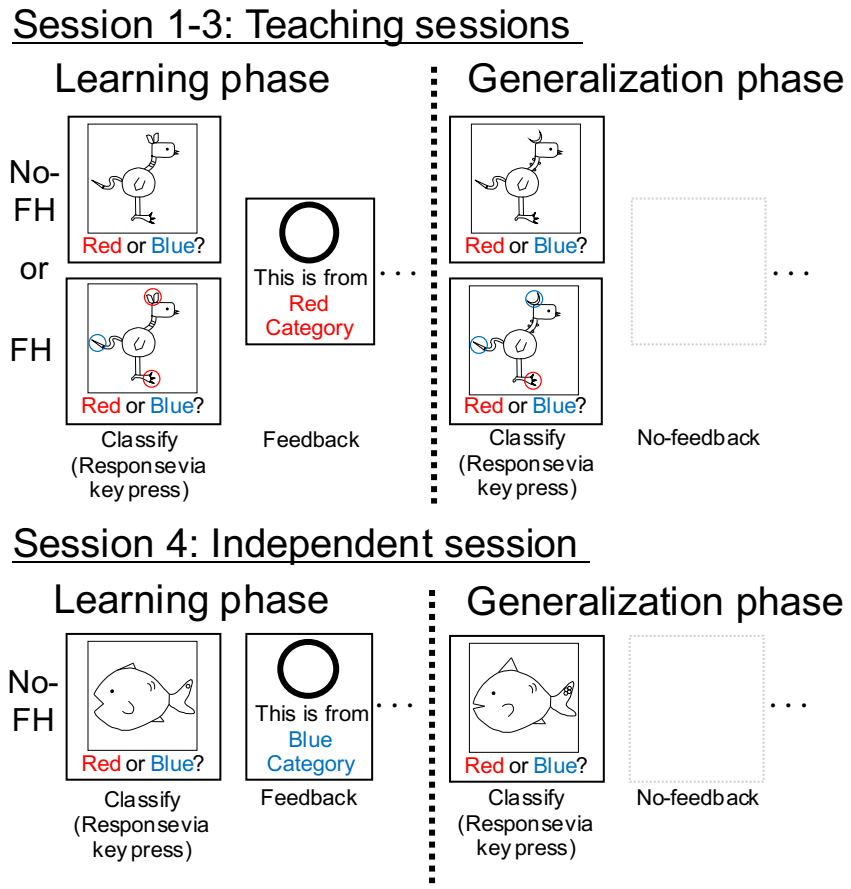
An Example of Stimuli Used in Experiments 1 and 2



Note. The features that are relevant to each category are circled with each category's color.

Figure 2

Procedure in Experiments 1and 2

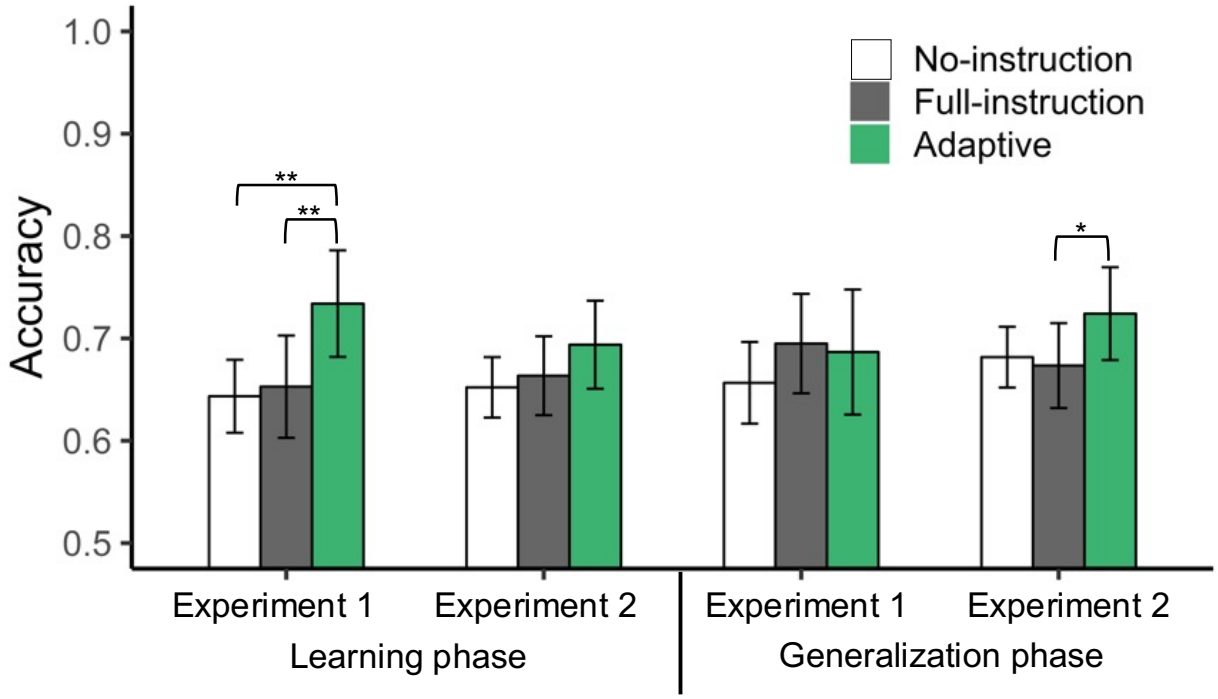


Note. On each trial, an item was shown and remained visible until response or 2000 ms. Participants categorized each item at their own pace by pressing the "f" or "j" key. In the teaching sessions (upper panels), feature highlighting was provided depending on the condition, highlighting relevant features with red or blue circles. In the independent session (lower panels), participants classified items from a new type of creature without feature highlighting. Each session included both the learning and generalization phases, except that in Experiment 1 the teaching sessions did not include the generalization phase. In the learning phase, participants classified each item, followed by feedback. In the generalization phase, participants classified

both the same items and novel items from the same creature, and feedback was not provided for these classifications.

Figure 3

Descriptive Statistics for Accuracy in the Learning and Generalization Phases



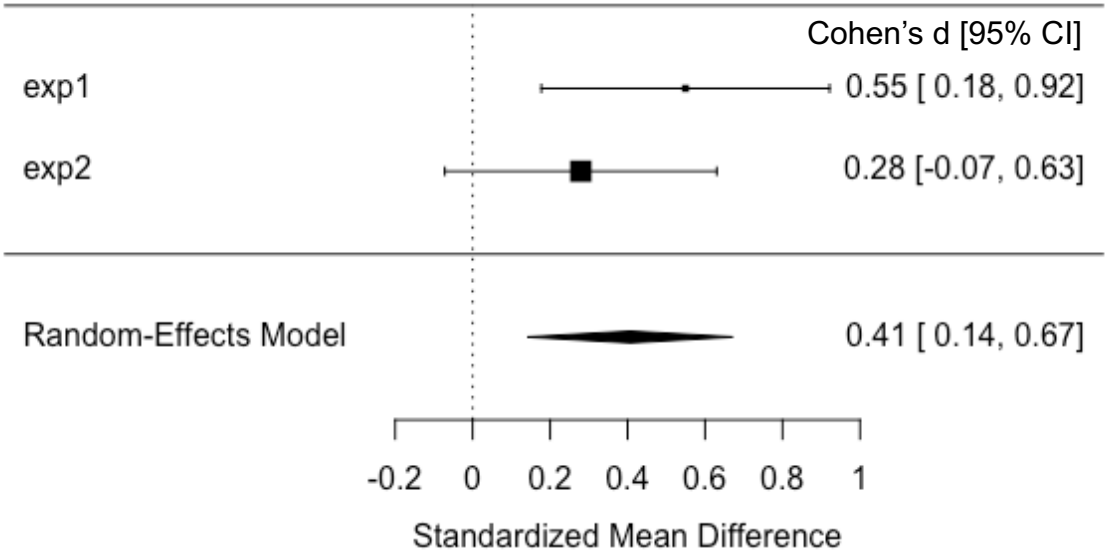
Note. Bars represent the mean accuracy in the learning and generalization phases across conditions. Error bars represent 95% confidence intervals. Asterisks indicate significant contrasts based on the regression analyses reported in Table S#.

$^{*}p < .05$. $^{**}p < .01$.

Figure 4

Results from the Meta-Analysis on Accuracy in the Learning Phase

Comparison between adaptive and no-instruction conditions



Comparison between adaptive and full-instruction conditions

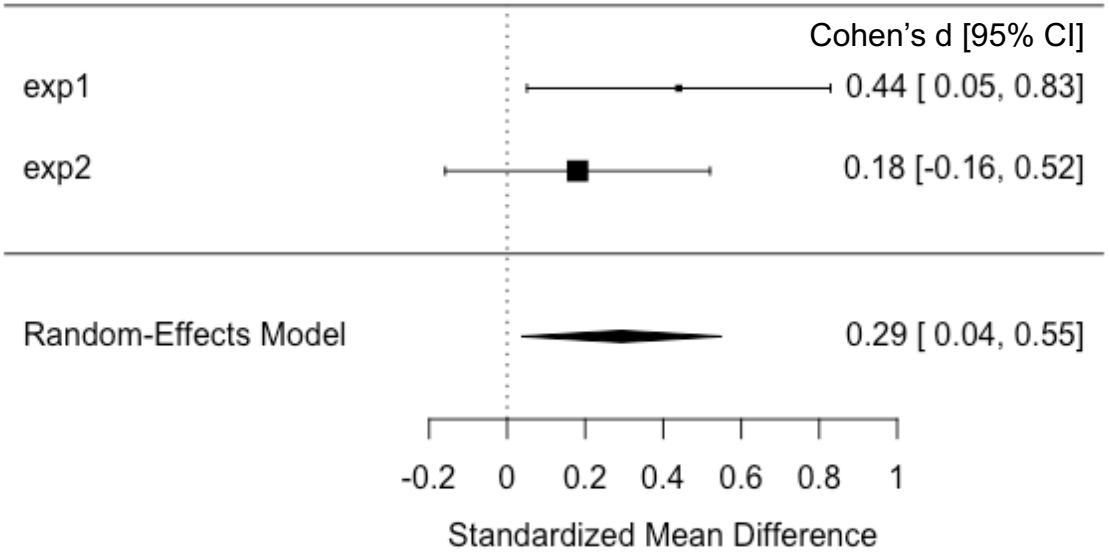
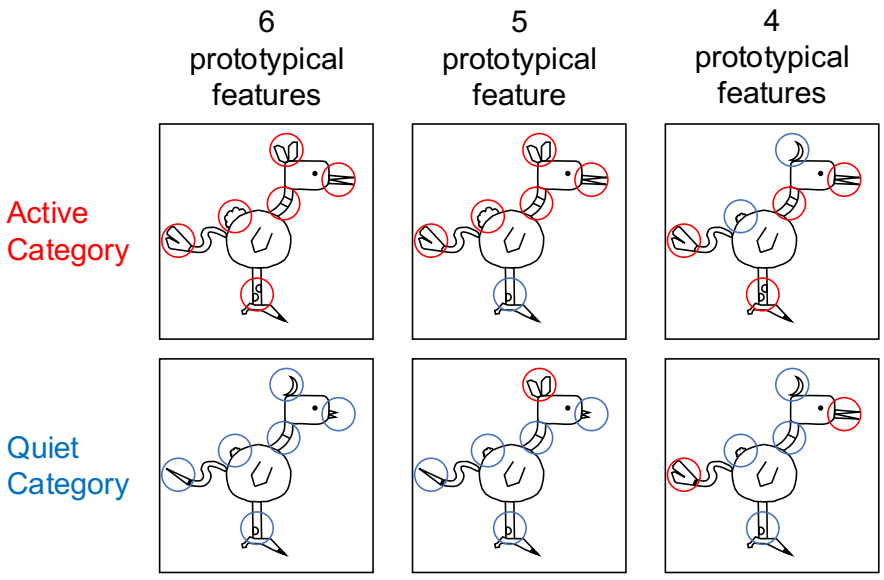


Figure 5

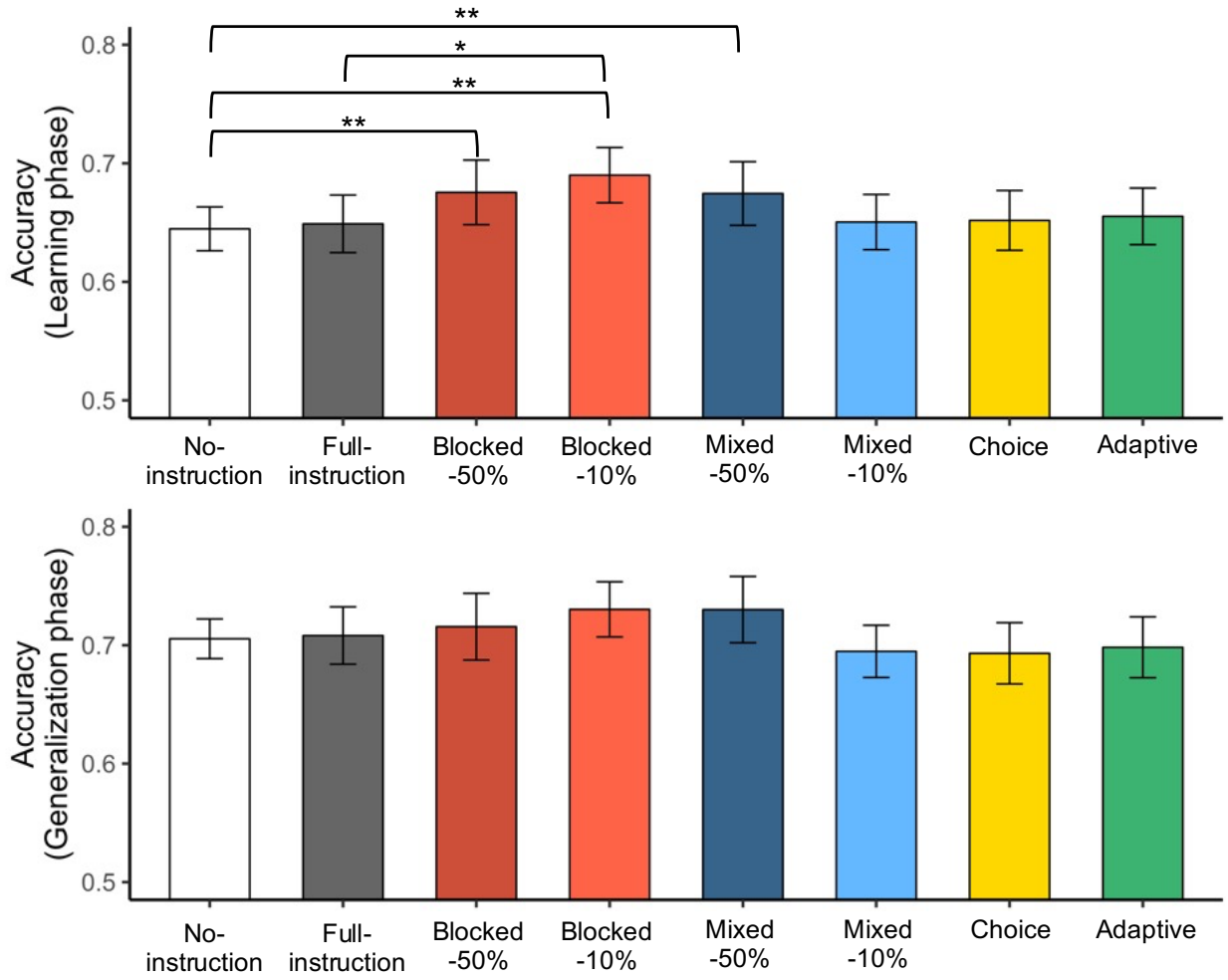
An Example of Stimuli Used in Experiment 3



Note. The features that are relevant to each category are circled with each category's color.

Figure 6

Descriptive Statistics for Accuracy in the Learning and Generalization Phases



Note. Bars represent the mean accuracy in the learning and generalization phases across conditions. Error bars represent 95% confidence intervals. Asterisks indicate significant contrasts based on the regression analyses reported in Table S#.

$^{*}p < .05$. $^{**}p < .01$.