Memory & Cognition (2021) 49:67–82
https://doi.org/10.3758/s13421-020-01074-w

Category similarity affects study choices in self-regulated learning

Xinyi Lu[1] · Trevor B. Penney[2] · Sean H. K. Kang[3]

Published online: 1 August 2020

Abstract

During learning, interleaving exemplars from different categories (e.g., ABCBCACAB) rather than blocking by category (e.g., AAABBBCCC) often enhances inductive learning, especially when the categories are highly similar. However, when allowed to select their own study schedules, learners overwhelmingly tend to block rather than interleave. Category similarity has been shown to moderate the relative benefit of interleaved versus blocked study. We investigated whether learners were sensitive to category similarity when choosing exemplars for study, and whether these choices predicted their learning outcomes. In Experiment 1, learners interleaved more often when the categories were highly similar (difficult to discriminate from each other), compared with when similarity was low. In Experiment 2, learners were presented with two sets of categories to learn; categories within each set were similar to each other, but categories were dissimilar across sets. When learners chose to interleave, they tended to switch to a similar rather than dissimilar category. Importantly, learners' study choices predicted their subsequent categorization performance. Our findings suggest that learners are strategic in their search for commonalities within versus differences among categories and can regulate their study behaviors based on category similarity.

Keywords Category learning · Self-regulated learning · Interleaved practice

When a learner is faced with a number of visual categories to learn, what is the optimal way to schedule practice? There are two broad ways a learner could sequence the category examples: One is to focus on learning one category at a time, repeatedly studying that category until it is mastered before moving on to the next category. This strategy is known as *blocked* practice (scheduling a single category's exemplars consecutively as a block). Another way is to alternate studying among the different categories, a strategy known as *interleaved* practice. In educational practice, blocking is far more common than interleaving, as evidenced by the huge majority of middle school math textbooks presenting problems in a blocked fashion (Rohrer, Dedrick, & Hartwig, in press). Indeed, most of the practice a learner encounters is likely to be blocked rather than interleaved; for example, a music student might focus on practicing a piece of music repeatedly until a certain degree of mastery has been attained before moving on to another piece (Maynard, 2006), while a mathematics teacher might complete classroom instruction and homework assignments for fractions before moving on to decimals (Rohrer, Dedrick, & Stershic, 2015).

However, a substantial body of research suggests that interleaved schedules are superior for many kinds of learning (Kang, 2017; Rohrer, Dedrick, Hartwig, & Cheung, 2019; Sana, Yan & Kim, 2017), including the learning of categories and concepts from examples. For example, Kornell and Bjork (2008) presented participants with paintings by 12 different artists, with each painting paired with the artist's last name (i.e., the category label) and either blocked or randomly interleaved by artist during study. Participants in the interleaved condition performed better than those in the blocked condition when asked to classify novel paintings by the studied artists. This interleaved over blocked performance advantage was found in other category learning studies, whether manipulated within or between subjects (e.g., Kornell, Castel, Eich, & Bjork, 2010; Verkoeijen & Bouwmeester, 2014).

✉ Sean H. K. Kang
sean.kang@unimelb.edu.au; seanhkkang@gmail.com

[1] Department of Psychology, University of Waterloo, Waterloo, ON, Canada

[2] Department of Psychology, The Chinese University of Hong Kong, Hong Kong, Hong Kong

[3] Melbourne Graduate School of Education, The University of Melbourne, Parkville, Australia

The role of category structure

According to the discriminative contrast hypothesis (Kornell & Bjork, 2008), interleaving category exemplars allows the juxtaposition of different categories, which supports the critical processes of category contrast and identification of the features that are unique to each category. This interpretation has found wide support; for example, Wahlheim, Dunlosky, and Jacoby (2011) had participants view single or paired exemplars from 12 bird families. Presenting exemplars in pairs improved performance when they were from different categories (i.e., interleaved), but not when from the same category (i.e., blocked). Moreover, when Birnbaum, Kornell, Bjork, and Bjork (2013) inserted 10 seconds of unrelated trivia questions between exemplar presentations of butterflies and birds, there was no interleaving benefit, suggesting that the trivia questions interfered with the ability to contrast exemplars of different categories (see also Kang & Pashler, 2012).

Carvalho and Goldstone (2014) argued for an extension to the discriminative contrast hypothesis. They postulated that while interleaving has been shown to be superior to blocking in most category learning studies, this result depends on the structure of the categories themselves. According to their attentional bias hypothesis, blocking allows one to notice the similarities between successive exemplars within a category. Conversely, interleaving facilitates noticing of the differences across categories. Therefore, one would expect blocked presentation to be superior when the exemplars within each category are highly variable (low similarity), and interleaved presentation to be superior when the categories are highly similar to each other. To test this hypothesis, they created blob-shaped categories of varying intercategory and intracategory similarity. In the low-similarity condition, the blob exemplars shared few similarities within and between categories, and they found that blocked presentation, which is supposed to promote commonality abstraction, produced better subsequent classification performance than did interleaved presentation. However, in the high-similarity condition, the blob exemplars shared a high level of similarity within and across categories, and they found that interleaved presentation was superior to blocked presentation, providing evidence that interleaving facilitates discrimination among the categories.

Overall, the literature shows that interleaved presentation tends to be superior to blocked presentation for learning natural categories, as interleaving allows for comparing and contrasting different categories and helps learners discover differences between categories. However, category structure is an important moderating factor (Brunmair & Richter, 2019; Carvalho & Goldstone, 2014). When intra-category variability is high, blocking helps learners discover the similarities within a category.

What do learners actually do during study?

While interleaved study tends to be superior for learning, researchers have discovered that learners hold a strong intuition for the opposite—that blocked study is more effective—even when their own test performance suggests otherwise (Kornell & Bjork, 2008; Yan, Bjork, & Bjork, 2016). When McCabe (2011) described the Kornell and Bjork (2008) experimental paradigm to undergraduates and asked them to predict which schedule would lead to better learning, an overwhelming majority predicted that blocked presentation would be better (only 6.67% correctly endorsed interleaved presentation). Another study that posed the same question to both college students and instructors found that only 16% and 13%, respectively, favored interleaving (Morehead, Rhodes, & Delozier, 2016).

In one of the few studies that explored participant choice in category learning directly (Tauber, Dunlosky, Rawson, Wahlheim, and Jacoby, 2013), participants could study either an exemplar from the same category as the current one or an exemplar from a different category (Experiments 1 & 2), or participants selected which category to study in every trial (Experiments 3 & 4). In all the experiments, a large majority of participants preferred to study the same category on subsequent trials rather than switch among categories.

Indeed, Yan et al. (2016) found that it was exceedingly difficult to uproot the metacognitive belief that blocked study is superior for learning. They proposed that learners persisted in their preference for blocking, even when their own performance in certain tasks suggested that interleaving was better, for several reasons: (1) blocked presentation feels subjectively easier, especially near the beginning of study, because of an increased sense of fluency, and (2) learners hold a priori beliefs that blocking should be more effective than interleaving, perhaps because blocked practice has been a common feature in their educational histories.

Although the emerging research on self-regulated category learning might suggest that learners are somewhat naïve when making study decisions during category learning, there is a substantial literature on metamemory demonstrating that learners can, in some situations, make relatively sophisticated study decisions that optimize memory performance. For instance, when choosing how to allocate study time among to-be-remembered items, learners make use of a combination of metacognitive monitoring of their learning of each item and their overall learning goals (e.g., Ariel & Dunlosky, 2013; Kornell & Metcalfe, 2006).

In another rare study that explored participant choice in category learning, Kornell and Vaughn (2018) found that participants who were allowed to choose which category they wished to study on every trial showed a preference for blocking that far exceeded chance levels. However, when participants *did* choose to switch away from a category, they tended not to return to it until they had studied a good number of other categories. This strategy was hypothesized to be a result of participants wishing to be thorough and fair in their category choices, as if they were "foraging" for information rather than choosing blocking or interleaving as a strategy per se. When asked to judge which kind of sequencing, pure blocking or pure interleaving, was more effective, the majority of participants chose pure blocking. However, when participants were asked to choose between pure blocking and a combination of interleaving and blocking, the majority chose the latter. This suggests that learners believe that a combination of both strategies is more effective than pure blocking, which is in turn believed to be more effective than pure interleaving.

The present study

While previous studies have shown that learner choice in category learning appears to be deliberate or systematic (e.g., Kornell & Vaughn, 2018; Tauber et al., 2013), there was no direct evidence that the behavior observed was strategic per se. In other words, the previous findings were purely descriptive accounts and were not a result of an experimental manipulation, and learner choice was not predictive of performance. In the current investigation, we manipulated category structure directly in a self-regulated learning paradigm to see whether this would cause participants' behavior to shift accordingly. Given that category structure moderates the relative benefit of interleaved versus blocked study for category learning (Brunmair & Richter, 2019; Carvalho & Goldstone, 2014), we wanted to investigate whether learners are sensitive to category similarity and adjust their study decisions in a way that promotes learning.

We decided to examine participants' study choices in a self-regulated paradigm, in which participants were free to choose which category to study at every trial. Our paradigm is similar to the Tauber et al. (2013) and Kornell and Vaughn (2018) studies in that we studied participants' choices directly, rather than asking them which strategy they preferred, and we allowed them to freely choose which category they wished to study next on each trial. However, our paradigm is different in important ways. First, we used fewer study categories (4–6 rather than 8–12), to reduce memory load, which may have biased participants toward blocking in earlier studies—for example, we can imagine that as the number of categories increases to eight, 10 or 12, participants may perceive the difficulty of interleaving to increase, as they must monitor all the categories they switch among, and may fall back to blocking, which seems cognitively simpler. Second, we examined whether participants' choices during study predicted their performance in a subsequent test. We predicted that although participants would show a preference for blocked study, the degree to which they blocked or interleaved would be strategic in nature. To study how participants' choices may be strategic and influenced by the task, in Experiment 1 we manipulated category similarity between participants by using the high-similarity and low-similarity blobs developed by Carvalho and Goldstone (2014). In Experiment 2, we manipulated category similarity within participants by using a selection of rock stimuli from Nosofsky, Sanders, Meagher, and Douglas (2017). All stimuli, data, and analysis files for both experiments are available (https://osf.io/6mrg2/).

Experiment 1

We hypothesized that when participants are allowed to sequence their own category learning, their study choices are strategic (i.e., they do not choose only a blocked strategy) and are influenced by category structure. Specifically, we expected participants to interleave more (i.e., block less) when the categories to be learned were high similarity, and to interleave less (i.e., block more) when the categories were low similarity. Furthermore, we expected participants' sequencing choices during study to predict their later categorization performance.

Method

Design We used a 2 (category similarity: high vs. low) × 3 (study sequence: self-regulated, blocked vs. interleaved) × 2 (test item: new vs. old) mixed-participant design, with category similarity and study sequence manipulated between subjects. The interleaved and blocked conditions were experimenter-controlled conditions that were aimed at replicating the interaction between category similarity and learning (Carvalho & Goldstone, 2014), as well as to provide a possible upper and lower limit to learning based on the most extreme forms of sequencing. Of primary interest were the self-regulated learning conditions in which participants could select which category to view on each study trial.

Materials The four categories to be learned were blobs developed by Carvalho and Goldstone (2014), with category membership defined by a specific notch in each blob. As shown in Fig. 1, there were two sets of blob categories: high similarity and low similarity. The high-similarity set contained exemplars that were highly similar within and across categories, whereas the low-similarity set contained exemplars that were low in similarity within and across categories.

Participants A total of 275 staff and students (ages 18–33 years, 175 females) from the Chinese University of Hong Kong were assigned to one of the six experimental conditions: (1) self-regulated high similarity (65 participants); (2) self-regulated low similarity (69 participants); (3) interleaved high

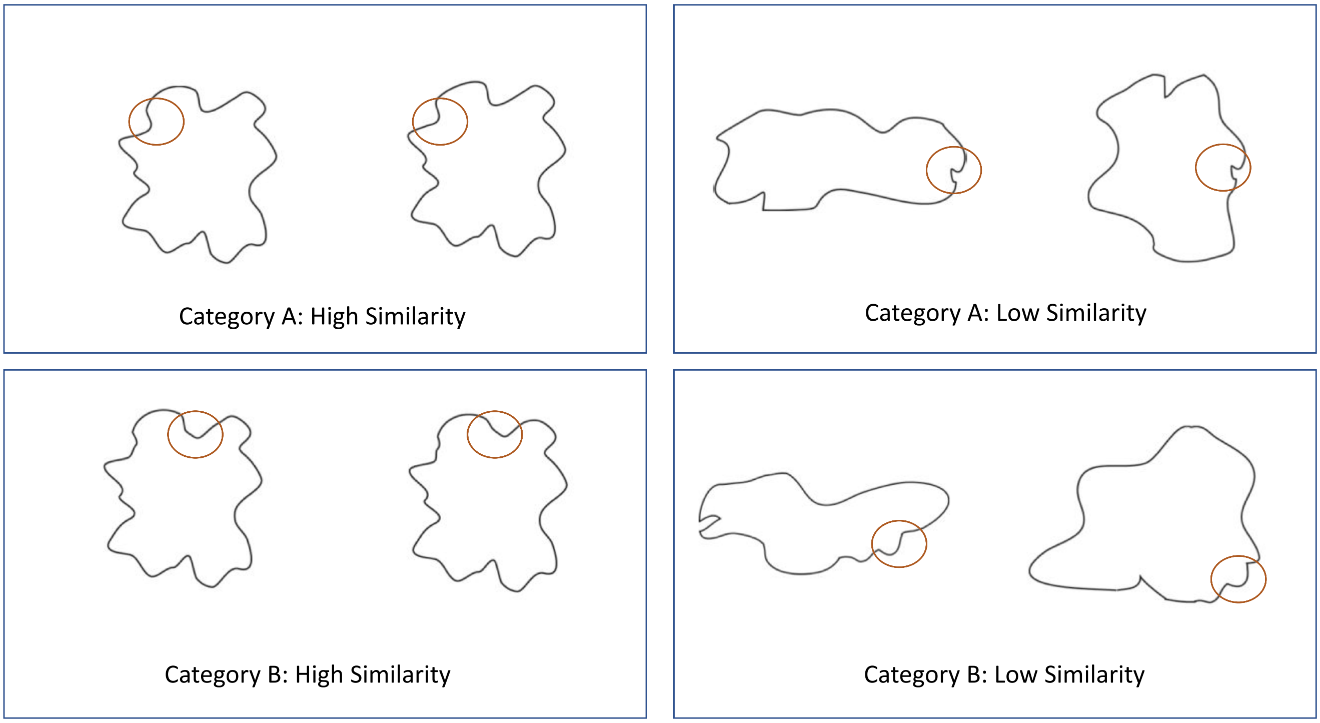
Category A: High Similarity
Category A: Low Similarity
Category B: High Similarity
Category B: Low Similarity

Fig. 1 Examples of blob stimuli used in Experiment 1. Category-defining features are circled. Blobs adopted from Carvalho and Goldstone (2014)

similarity (33 participants); (4) interleaved low similarity (37 participants); (5) blocked high similarity (39 participants); (6) blocked low similarity (32 participants). We set an *a priori* minimum sample size of 30 participants per cell based on Experiment 1 (n = 31 and 29) of Carvalho and Goldstone (2014); this gave us an estimated .84 power (predicted d = 0.76 for effect of interleaving). However, we did not have an *a priori* estimate on the effect of similarity on study behavior, which made estimating required sample size for this analysis difficult. Therefore, we elected to run post hoc sensitivity analyses based on simulations, which we report in our results (Green & MacLeod, 2016; Johnson, Barry, Ferguson, & Müller, 2015). We also made the decision that in order to obtain more power for analyzing our primary effect of interest (the self-regulated study behaviors), we would double the probability that participants would be allocated to a self-regulated condition.

Procedure Participants were assigned to one of the six experimental conditions. The task scenario given to participants was that the blobs were newly discovered alien cell species that they had to learn to distinguish. The four "species" were labeled P, Q, R, and U, with the assignment of letter labels to categories randomized for each participant. There were 16 exemplars in each of the four categories, of which eight were randomly selected to appear in the learning session. These learning phase exemplars each appeared six times to make a total of 192 study trials (8 × 4 × 6) in the learning session.

For participants in the self-regulated conditions, each learning trial began with the choice screen that had four clickable category selection buttons, along with text below that asked which category they wished to see next. The order of the buttons was randomized for each participant. Each category selection button showed the name of the category and the number of exemplars that remained available for study in parentheses. Once a participant exhausted all 48 exemplars for a category (i.e., the number available for study reached zero), the category button would remain on the screen, but would not be clickable by the participant. Upon the participant's selection, a fixation cross appeared in the center of the screen for 500 ms, followed by a category exemplar and its category label for 2,000 ms (see Fig. 2).

For participants in the experimenter-controlled conditions, each learning trial began with a fixation cross for 500 ms, followed by a category exemplar and its category label for 2,000 ms, followed by a button that asked them to click next to proceed to the next item.

After the learning phase, participants were asked a series of demographic questions, and then were given a series of five simple arithmetic questions (e.g. "What is 10 + 12?") as a brief filler task. Participants took 32.5 s on average to complete this task. The test phase began immediately after. For the classification test, all 16 exemplars from each category were included such that half of the test items were novel and half of the items were studied before, for a total of 64 test items. The sequence of test items was randomized for each participant.

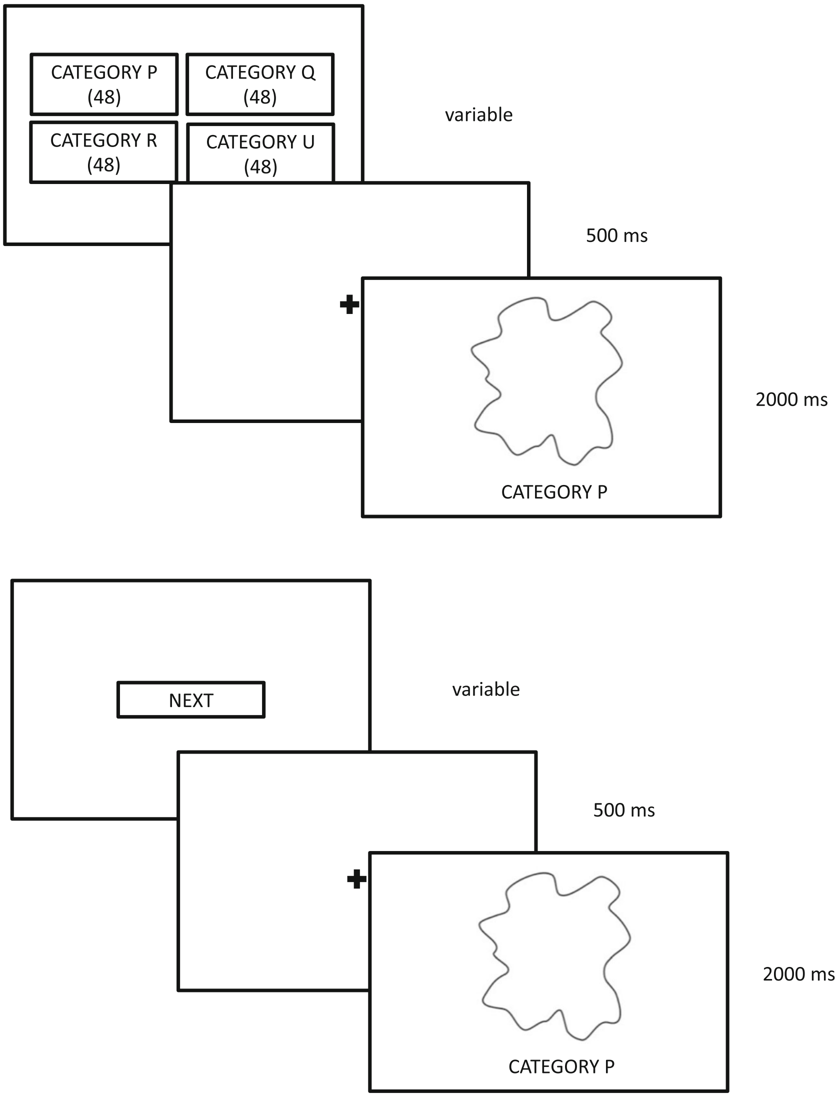

Fig. 2 Sequence of events during a learning trial for the self-regulated conditions (top) and experimenter-controlled conditions (bottom) in Experiment 1

Results and discussion

We first investigated performance in the experimenter-controlled conditions (blocked and interleaved) separately from the self-regulated conditions. Aside from the self-regulated versus experimenter-controlled factor, other between-participant factors were category similarity (high vs. low) and study sequence (blocked vs. interleaved); also, whether a test item was old or new was analyzed as a within-participant factor.

Performance in experimenter-controlled blocked and interleaved conditions A 2 (category similarity: high vs. low) × 2 (study sequence: blocked vs. interleaved) × 2 (test item: new vs. old) mixed-factors analysis of variance (ANOVA) was used to analyze the categorization performance data in the control conditions. Performance was better for participants learning the low-similarity categories compared with the high-similarity categories, $F(1,\ 137) = 5.85$, $p = .017$, $\eta^2_G = 0.038$, as well as for old items compared with new items, $F(1,\ 137) = 32.1$, $p < .001$, $\eta^2_G = 0.018$. This was qualified by a significant interaction between similarity and old/new, $F(1,\ 137) = 20.5$, $p < .001$, $\eta^2_G = 0.012$, which manifested in an old item over new item advantage for learners studying the low-similarity categories, $t(68) = 6.01$, $p < .001$, $d = 0.72$, but not the high-similarity categories, $t(71) = 1.14$, $p = .26$, $d = 0.13$. Critically, although the effect of study sequence was not significant, $F(1,\ 137) < 1$, there was a significant interaction between study sequence and similarity, $F(1,\ 137) = 13.76$, $p < .001$, $\eta^2_G = 0.090$, as shown in Fig. 3. When similarity was high, interleaved study was superior to blocked study, $t(142) = 6.86$, $p < .001$, $d = 1.15$, but when similarity was low, the difference was not significant, $t(136) = 1.21$, $p = .23$, $d = 0.21$. The three-way interaction was not significant, $F(1,\ 137) < 1$.

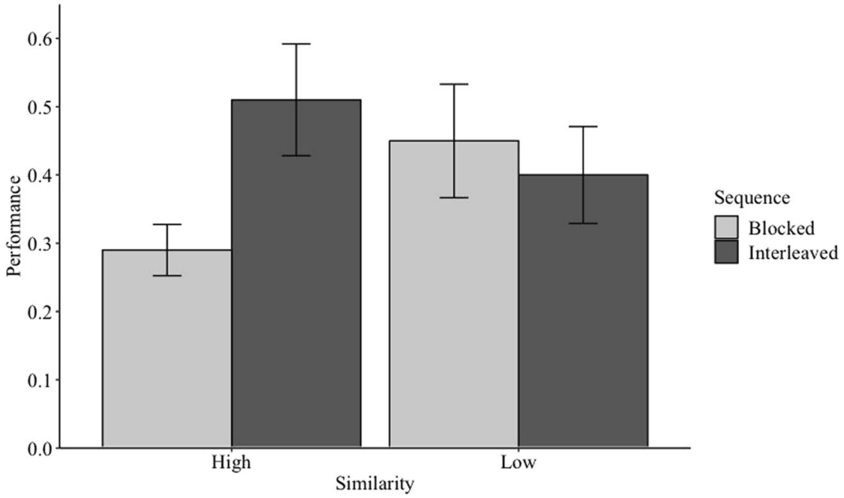

Fig. 3 Proportion correct as a function of category similarity and study sequence. Error bars show standard errors

Comparing performance in self-regulated to experimenter-controlled conditions To compare the self-regulated conditions to the control conditions, we collapsed across blocked and interleaved conditions and performed a 2 (category similarity: high vs. low) × 2 (study sequence: self-regulated vs. control) × 2 (test item: new vs. old) mixed-factors ANOVA. There was a main effect of sequence, $F(1, 271) = 35.1$, $p < .001$, $\eta^2_G = 0.109$, indicating that participants who self-regulated their learning ($M = 0.59$, $SD = 0.28$) performed better than when sequence was experimenter controlled ($M = 0.41$, $SD = 0.23$). There was also main effect of similarity. $F(1, 271) = 5.68$, $p = .018$, $\eta^2_G = 0.019$, with low-similarity categories better categorized than high-similarity categories, and a main effect of old/new, $F(1, 137) = 77.7$, $p < .001$, $\eta^2_G = 0.015$, with old items categorized better than new items. The three-way interaction was not significant, $F(1, 271) = 1.01$, though there was again a significant interaction between similarity and old/new, $F(1, 137) = 28.9$, $p < .001$, $\eta^2_G = 0.006$. Table 1 summarizes the classification performance of the six different study conditions.

The results of the experimenter-controlled conditions replicated Carvalho and Goldstone's (2014) main finding that category similarity moderates whether an interleaved or blocked schedule is superior for category learning. When the categories to be learned were highly similar to one another and thus difficult to tell apart, the discriminative contrast that interleaved presentation facilitated was most beneficial. However, when the exemplars within a category were quite dissimilar to each other and thus difficult to reconcile as being from the same category, an interleaved schedule was no longer optimal. However, unlike Carvalho and Goldstone (2014), we did not find that a blocked schedule significantly benefited category learning in the low-similarity condition (although there was a trend in that direction); this could be because our sample size was smaller and/or our task parameters were slightly different from theirs. We also found a significant similarity by old/new interaction. The interaction appears to be driven by a high-similarity (M = .47, SD = 0.25) versus low-similarity (M = .58, SD = 0.28) difference in the old test items, but not the new test items (M = .45, SD = 0.26 vs. M = .48, SD = 0.28). This may be because the low-similarity items are more distinct than the high-similarity items (see Fig. 1), having more unique shapes and features, and are therefore more memorable. Participants could then rely on these memories when classifying the old items, but not the new items.

Table 1 Mean proportion correct of the different study conditions in Experiment 1 (standard deviations shown in parentheses)

Condition	Old items	Novel items
Self-regulated high similarity	0.55 (0.27)	0.51 (0.28)
Self-regulated low similarity	0.69 (0.27)	0.59 (0.29)
Blocked high similarity	0.31 (0.13)	0.28 (0.13)
Blocked low similarity	0.51 (0.27)	0.40 (0.23)
Interleaved high similarity	0.51 (0.26)	0.51 (0.24)
Interleaved low similarity	0.45 (0.24)	0.35 (0.21)

An important novel finding is that the self-regulated participants performed better than participants in the experimenter-controlled conditions. The act of choosing which categories to study (compared with just passively receiving the information) may have increased motivation and engagement in the learning task.

Self-regulated study behavior

We analyzed participants' self-regulated study behavior based on our main measure of choice: the proportion of switching between different categories ("proportion of switching"). As the experiment had 192 learning trials, each participant could make 191 choices, and each choice could be classified as a "switch" or a "stay." Therefore, the proportion of switching was calculated by dividing the number of times a participant chose to switch by the total number of possible choices (191). This yielded a number between 0.016 (which would be considered pure blocking, apart from three obligatory category switches) and 1 (which would be considered pure interleaving). A high proportion of switching indicates a tendency toward interleaving (choosing to switch between different categories), whereas a low proportion of switching indicates a tendency toward blocking (choosing to repeat the same category). A Welch two-sample *t* test, $t(127.05) = 9.49$, $p < .001$, $d = 1.65$, indicated that participants studying the high-similarity categories tended to switch much more often (64.38%) compared with participants studying the low-similarity categories (28.29%).

We performed a generalized linear mixed-effects analysis of the effect of category similarity and proportion of switching during study on subsequent categorization performance using the lme4 package in R (Bates, Mächler, Bolker, & Walker, 2015). Fixed effects entered into the model were category similarity and proportion of switching, and their interaction term, as well as whether an item was new or old. Intercepts for subjects and items were included as random effects. A parallel linear regression analysis revealed similar results. The final model parameters (Akaike information criterion [AIC] = 8,930.5] are reported in Table 2.

The results indicate that new items were associated with worse performance (odds ratio = 0.65) compared with old (studied) items, and high-similarity categories were associated with worse performance (odds ratio = 0.50) compared with low-similarity categories, which were in line with our expectations. However, the proportion of switching did not significantly predict categorization performance, nor was the interaction between category similarity and the proportion of switching significant. We had predicted that, for participants learning the high-similarity categories, switching more during study (i.e., a tendency to interleave) would enhance their subsequent categorization performance. Conversely, we predicted that for participants learning the low-similarity categories, switching less during study (i.e., a tendency to block) would enhance their subsequent categorization performance. Although there were trends in the hypothesized direction, these did not reach significance. A post hoc sensitivity analysis with 1,000 replications conducted using the *simr* package in R (Brysbaert & Stevens, 2018; Green and MacLeod, 2016) indicated that we had .50 power to detect the effect of similarity, .25 power to detect the effect of proportion of switches, and .12 power to detect their interaction term. We also noticed that the confidence intervals for the predicted odds ratios were very wide, indicating very high variability in the data. Plotting performance as a function of similarity and proportion of switching revealed an apparent trend for switching to improve performance in the high-similarity condition (i.e., a steeper slope), but not in the low-similarity condition (i.e., a flatter slope; see Fig. 4).

Table 2 Mixed-effects model components

Predictors	Correct response		
	Odds ratios	CI	p
(Intercept)	1.62	0.80, 3.31	.183
New vs. old	0.65	0.59, 0.73	**<.001**
Similarity: High vs. low	0.50	0.25, 1.02	.057
Proportion of switching	2.00	0.52, 7.75	.315
Similarity × Proportion of Switching	1.62	0.42, 6.26	.486
Random effects			
σ^2	3.29		
$\tau_{00\ \text{subject\_id}}$	2.83		
$\tau_{00\ \text{stimulus}}$	0.22		
ICC $_{\text{subject\_id}}$	0.45		
ICC $_{\text{stimulus}}$	0.03		
Observations	8,576		
Marginal R^2/Conditional R^2	0.030/0.497		

Bold indicates $p < .05$

Exploratory analyses

Aside from the proportion of switching, we also examined three additional measures in the self-regulated study choices, summarized in Table 3: (1) the length of each participant's longest blocked sequence, defined as a stretch of choices where a single category was repeatedly chosen; (2) the average length of a participant's blocked sequences; (3) the length of each participant's longest interleaved sequence, defined as a stretch of choices where up to one consecutive category repeat could occur (e.g., ABCABCCABC). In all the measures we examined, there was a significant difference between participants studying the high-similarity categories versus the low-similarity categories. Of course, these different measures will all be correlated with each other to some extent, as they are different ways of measuring the same underlying construct (i.e., the degree of blocking vs. interleaving).

We also examined whether the tendency to block or interleave shifted across the duration of the experiment. We split each participant's choices into four quarters of 48 choices each (except the first quarter, which had 47 choices) and calculated the proportion of switches for each quarter. A mixed-factor ANOVA was used to compare the proportion of switching across quarter as a function of category similarity. There was a main effect of similarity, $F(1, 13) = 90.88$, $p < .001$, $\eta^2_G = .316$, showing that participants switched more often in the high-similarity condition overall. There was also a main effect of quarter, $F(2.31, 304.32) = 33.64$, $p < .001$, $\eta^2_G = .077$,

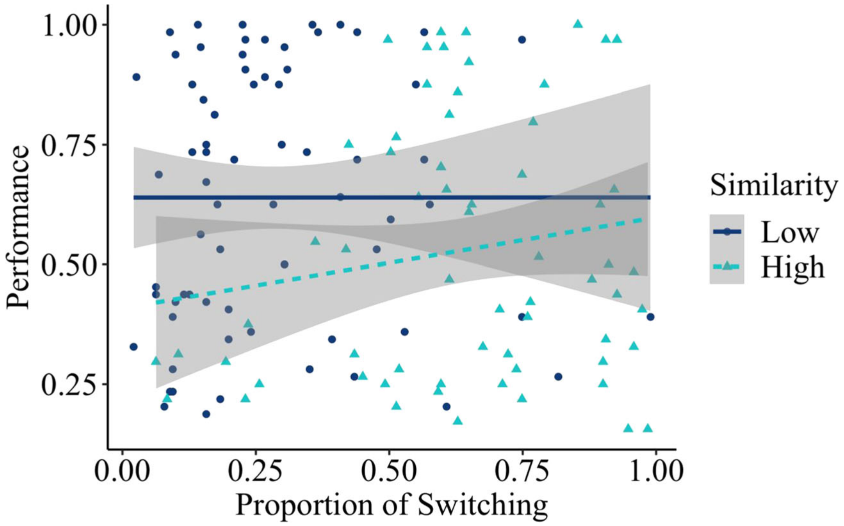

Fig. 4 Categorization performance as a function of the proportion of among-category switching ("proportion of switching"). Shaded areas represent 95% confidence intervals

suggesting that there were differences in the tendency to interleave across quarters. Since the interaction between similarity and quarter was significant, $F(2.31, 302.32) = 8.94$, $p < .001$, $\eta^2_G = .022$, we followed up with a separate one-way ANOVA for the high and low similarity groups (see Fig. 5).

In the high-similarity group, there was a significant effect of quarter, $F(2.17, 138.76) = 22.27$, $p < .001$, $\eta^2_G = .109$. Bonferroni-corrected pairwise *t* tests revealed that there was a higher proportion of switching in the first quarter compared with the remaining three quarters, and a higher proportion of switching in the second quarter compared with the last two quarters. This suggests that the proportion of switching was highest at the start of learning and decreased significantly across the duration of the study phase. In the low-similarity group, there was a significant effect of quarter, $F(2.31, 157.12) = 18.72$, $p < .001$, $\eta^2_G = .076$. Bonferroni-corrected pairwise *t* tests revealed that there was higher proportion of switching in the first quarter compared with the other three quarters, suggesting that the proportion of switching was highest at the start of learning, but leveled off after the first quarter. Although exploratory, these results would appear to suggest that learners interleave more at the start of learning rather than toward the end.

Experiment 2

Experiment 1 manipulated category similarity between participants and showed that learners do regulate their study behaviors according to the material they are presented with. However, our prediction that learner's study choices would affect their subsequent performance was not borne out by the data, though there was a weak trend in line with our hypotheses (shown in Fig. 4).

We were therefore motivated to conduct Experiment 2 as an extension of Experiment 1. Our goals were to replicate the finding that similarity motivates category switching behavior, as well as examine whether these changes in behavior predict performance. In Experiment 1, although there was a robust effect of category similarity on study behavior, there was no effect of choice in predicting learner performance. We decided to use a within-participants design in Experiment 2 as a way to potentially increase the salience of category similarity (each participant experienced both high-similarity and low-similarity categories) and also reduce the variability in the data.

Naturalistic rock categories were chosen as stimuli rather than the blobs used in Experiment 1. The reasoning for this

Table 3 Mean measures of the degree of participants' tendency toward blocking and interleaving

Mean measure	Condition		*p*	Cohen's *d*
	High similarity	Low similarity		
Proportion of switching	.64	.28	<.001	1.65
Longest blocked sequence length	14.4	23.5	<.001	0.72
Average blocked sequence length	6.3	11.6	<.001	0.66
Longest interleaved sequence length	74.3	23.7	<.001	1.13

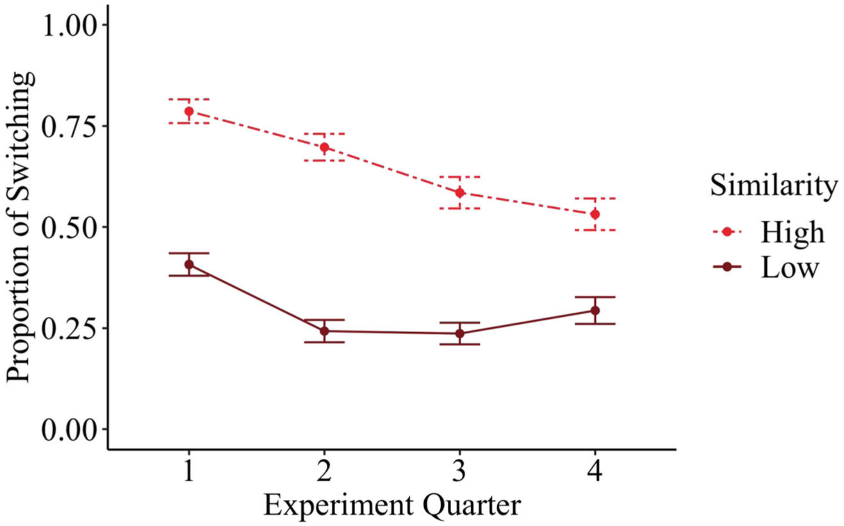

Fig. 5 Proportion of switching as a function of experiment quarter and category similarity. Error bars show standard errors

was manifold. First, the blob categories were defined by a single identifying feature (by design), whereas most previous research on interleaving versus blocking involved naturalistic categories, such as paintings. While artificial categories tend to be defined by a set of easily verbalizable rules ("rule-based" categories), naturalistic categories are usually much more fuzzily defined by a combination of characteristics ("information integration" categories; Ashby & Maddox, 2011). Therefore, to increase the generalizability of our findings, we used rock categories that are similar to the naturalistic categories used in many previous interleaving experiments (and which are more in line with what a learner would expect to encounter in the real world). Also, as we wished to compare similar and dissimilar categories within participants, the blob categories would prove to be problematic for this purpose as the degree of similarity is confounded with difficulty—the highly similar categories are harder to learn than the dissimilar categories. Finally, many participants reported finding the blobs themselves extremely difficult to learn (and even to tell apart). The difficulty of the learning task meant that the non-significant effect of switching on performance was difficult to interpret; it may well be that many more learning trials are required to see a shift in test performance. In conducting Experiment 2, we hoped to resolve these issues by using naturalistic categories that were easier to learn and required fewer stimulus repetitions.

In Experiment 2, learners were presented with two sets of rock categories to learn on every trial; categories within each set were similar to each other, but categories were dissimilar across sets. We predicted that participants would be sensitive to which categories were similar and which were dissimilar and adjust their study strategy accordingly. Our main prediction was that when learners do choose to interleave, they will tend to switch among categories that are similar to each other more than among categories that are dissimilar. Furthermore, we expected participants' sequencing choices during study to predict their later categorization performance.

Method

Materials We used two sets of three categories each. Within each set, the three categories were highly similar to each other, while being dissimilar to the categories in the other set.

We selected a subset of the normed rock stimuli from Nosofsky et al. (2017) that fulfilled the above requirements for category similarity. Each category consisted of 12 unique exemplars, of which half were randomly selected to appear in the learning phase. Table 4 shows the average pairwise similarity ratings for the rock categories as reported by Nosofsky et al. (2017). Anthracite, bituminous coal, and obsidian (average within-set similarity: 6.45) formed one set of three, whereas dolomite, quartzite, and micrite formed another set of three (average within-set similarity: 5.20), and the sets were dissimilar to each other (average cross-set similarity: 3.50). Examples of the six categories are shown in Fig. 6.

Participants and procedure A total of 84 undergraduates from Dartmouth College participated in the experiment. We set a minimum sample size of 80 based on an *a priori* power analysis (.80 power to detect the effect of proportion of switching on performance using the predicted odds ratio from Experiment 1). We did not include any blocked or interleaved control conditions, but focused solely on self-regulated participants with category similarity as a within-participants manipulation. The procedure during the learning phase was similar to the self-regulated conditions in Experiment 1, with participants choosing which exemplar to study next (see Fig. 7). On every trial, six clickable category selection buttons appeared, and clicking a button brought up the image of a category exemplar. The order of buttons was again randomized for each

Table 4 Average pairwise similarity ratings for the rock categories as reported by Nosofsky et al. (2017). Diagonal indicates intracategory similarity

	Anthracite	Bituminous coal	Obsidian	Dolomite	Quartzite	Micrite
Anthracite	7.35					
Bituminous coal	6.93	7.11				
Obsidian	6.47	5.96	7.15			
Dolomite	3.26	3.78	2.92	4.95		
Quartzite	3.63	3.74	3.21	4.86	5.85	
Micrite	3.72	4.21	3.11	5.45	5.29	5.80

participant; therefore, similar categories were not systematically grouped together. Each unique category exemplar appeared twice in the learning phase, for a total of 12 trials per category, and 72 trials in total. (As pilot data indicated that the rocks were easier to learn than the blobs used in Experiment 1, we reduced the number of stimulus repetitions to avoid a ceiling effect.)

After the learning phase, participants were asked a series of demographic questions, and then were given a series of five simple arithmetic questions as a brief filler task. Participants took 36.3 s on average to complete this task. The test phase began immediately after. For the classification test, all 12 exemplars from each category were included such that half of the test items were novel and half of the items were studied before, for a total of 72 test items. The sequence of test items was randomized for each participant.

Results and discussion

For the final test, mean categorization performance for the previously studied category exemplars was .54 (*SD* = .12), while for the new exemplars it was .42 (*SD* = .11).

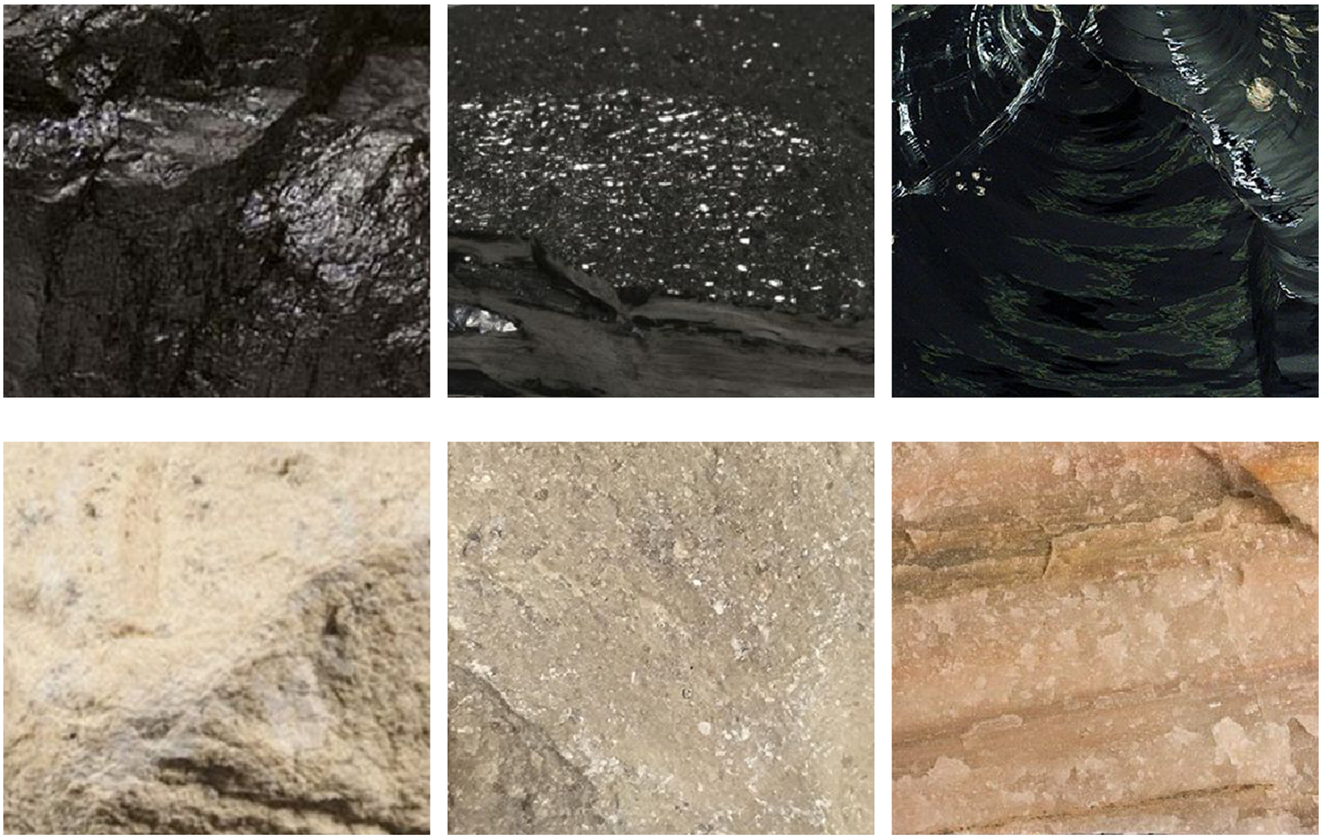

Fig. 6 Examples of rock stimuli used in Experiment 2. Top row (left–right): anthracite, bituminous coal, obsidian. Bottom row (left–right): dolomite, micrite, quartzite. Stimuli from Nosofsky et al. (2017)

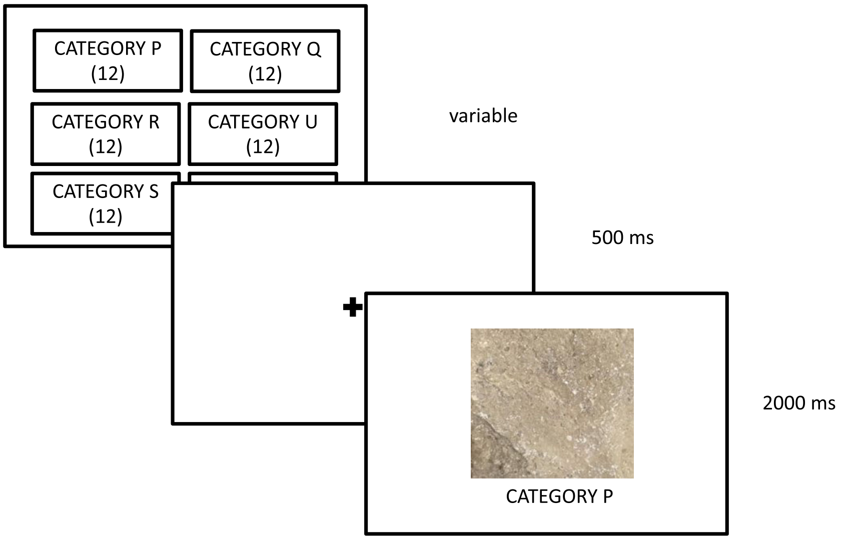

Fig. 7 Sequence of events for a learning trial in Experiment 2

Considering that baseline performance by chance would be 1/6, compared with a baseline of 1/4 in Experiment 1, it appeared that the rock categories were indeed easier to learn with fewer repetitions than the blob categories.

As in Experiment 1, we calculated the proportion of switching among different categories ("proportion of switching") by dividing the number of switch choices by the total number of choices. The proportion of switching was .48, far below a chance level of 5/6 or .83, which was suggestive of participants' preference for blocked study. We also calculated the proportion of switching among between similar rather than dissimilar categories ("similar switch ratio") by dividing each participant's number of similar switches by their total number of switches (see Table 5). If a participant chose to switch to a different category, he or she could switch to either a similar category (of which there were two) or a dissimilar category (of which there were three). Therefore, if a participant's choices were completely nonstrategic and based on chance alone, we would expect the similar switch ratio to be around 2/5, or 0.40. However, a one-sample *t* test, $t(83) = 4.81$, $p < .001$, $d = 0.52$, revealed that the mean similar switch ratio of 0.51, 95% CI [0.46, 0.55] was significantly higher than 0.40. This result indicates that participants were much more likely to interleave among similar categories than the chance baseline.

Table 5 Measures of the degree of participants' tendency to switch among categories

Measure	Mean
Number of switches	34.3
Number of similar-category switches	18.9
Proportion of switching	.48
Similar switch ratio	.51

Next, we examined whether these study choices predicted performance. We performed a generalized linear mixed-effects analysis of the effects of proportion of switching and the similar switch ratio on subsequent categorization performance. Fixed effects entered into the model were proportion of switching and similar switch ratio, and their interaction term, as well as whether an item was new or old. Intercepts for participants and items were included as random effects. A parallel linear regression analysis revealed similar results. The final model parameters [AIC = 7,868.9] are reported in Table 6.

The results indicate that new items were associated with worse performance (odds ratio = 0.77) compared with old (studied) items, which is in line with our expectations. As in Experiment 1, we used the proportion of switching to predict how well participants did in the subsequent categorization task. Comparing the odds ratios for the effect of switching in Experiment 1 (2.00) and Experiment 2 (3.92), the effect of switching was larger in Experiment 2. We found that the trends we observed in Experiment 1 were now statistically significant: the proportion of between-category switching significantly predicted categorization performance (i.e., more interleaving was associated with enhanced category learning; see Fig. 8).

Importantly for our hypothesis, the similar switch ratio also significantly predicted performance. As Fig. 9 shows, when participants chose to interleave, switching among similar items rather than dissimilar items during study enhanced participants' category learning.

A post hoc sensitivity analysis conducted with 1,000 replications indicated that we had .66 power to detect the effect of proportion of switches, .75 power to detect the effect of the similar switch ratio, and .48 power to detect their interaction term.

Table 6 Mixed-model components

Predictors	Correct response		
	Odds Ratios	CI	p
(Intercept)	0.39	0.23, 0.66	**.001**
New vs. old	0.77	0.72, 0.81	**<.001**
Similar switch ratio	4.25	1.41, 12.82	**.010**
Proportion of switching	3.92	1.23, 12.44	**.020**
Similar Switch Ratio × Proportion of Switching	0.13	0.02, 1.15	.066
Random effects			
σ^2	3.29		
$\tau_{00\ \text{subject\_id}}$	0.12		
$\tau_{00\ \text{stimulus}}$	0.38		
ICC	0.13		
$N_{\text{subject\_id}}$	84		
N_{stimulus}	72		
Observations	6,048		
Marginal R^2/Conditional R^2	0.027/0.155		

Bold indicates $p < .05$

Exploratory analyses

We again examined whether the tendency to block or interleave shifted across the duration of the experiment. We split each participant's choices into four quarters of 18 choices each (except the first quarter, which had 17 choices) and calculated both the proportion of switching and the similar switch ratio for each quarter (see Table 7). A pair of one-way repeated-measures ANOVAs with Greenhouse–Geisser corrections showed that both the proportion of switching, $F(2.32, 192.50) = 3.88$, $p = .017$, $\eta^2_G = .014$, and the similar switch ratio, $F(2.82, 234.18) = 7.34$, $p < .001$, $\eta^2_G = 037$, were significantly different between quarters. Pairwise t tests (Bonferroni corrected) revealed that there was a higher proportion of switching in the first quarter than in the second quarter, and that the similar switch ratio was higher in the third quarter compared with the first quarter, as well as the third and second quarters compared with the fourth quarter.

Although exploratory, these results would appear to suggest that learners were interleaving more during the first block in a process of exploration and discovery. Since learners were not told of the existence of similar–dissimilar category sets, they had to discover this structure on their own. Also, the increased similar switch ratio in the middle two blocks would

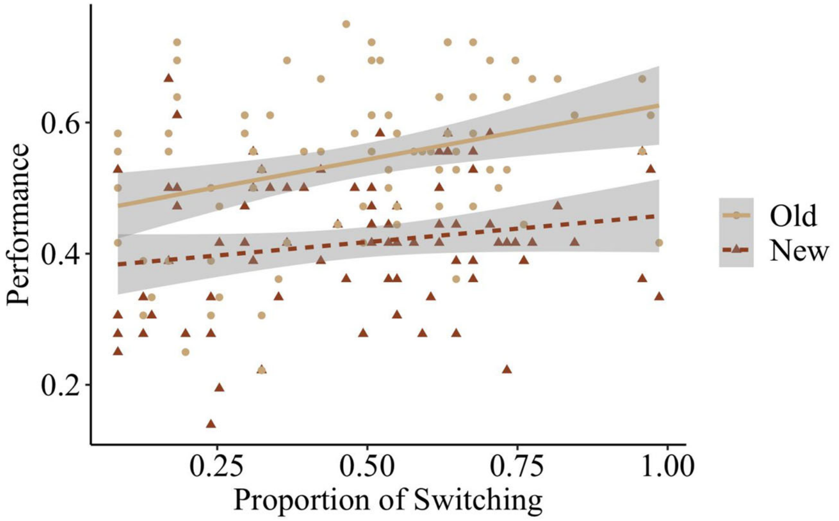

Fig. 8 Categorization performance as a function of the proportion of among-category switching ("proportion of switching"). Shaded areas represent 95% confidence intervals

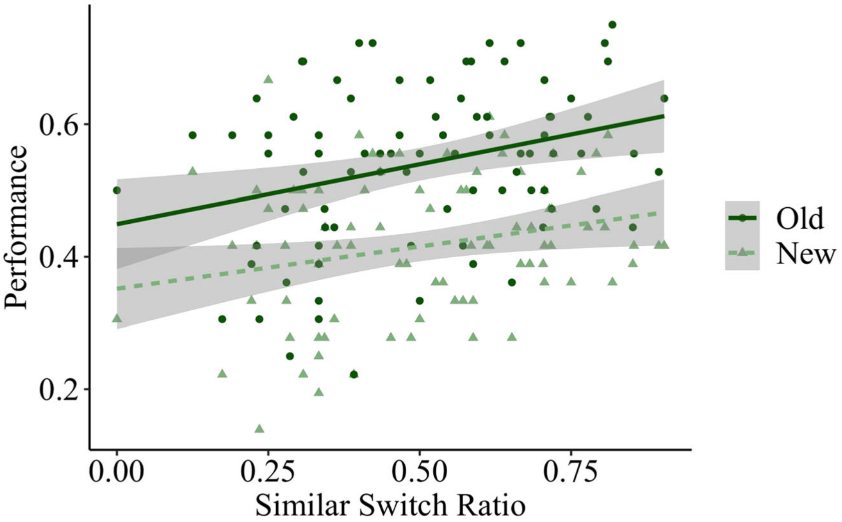

Fig. 9 Categorization performance as a function of the proportion of switching among similar categories ("similar switch ratio"). Shaded areas represent 95% confidence intervals

suggest that, upon discovery of this structure, they focused their switching to be among similar categories.

General discussion

Across two experiments, we found that participant's study choices during category learning were influenced by the similarity of the to-be-learned categories. In Experiment 1, participants who were learning highly similar categories tended to interleave the exemplars more often during study compared with participants who were learning categories that were low in similarity. Furthermore, in the high-similarity condition, we noted that increased interleaving appeared to be associated with better performance in a later categorization test, but not in the low-similarity condition, although this trend did not reach significance. Experiment 2 showed that when participants were learning both similar and dissimilar categories (within-participants design), they tended to interleave the similar categories more than they interleaved the dissimilar categories, which suggests that learners have some awareness of how interleaving might be more beneficial for those categories (as in the discriminative contrast hypothesis; e.g., Kang & Pashler, 2012). Interleaving among similar categories helps learners discover the subtle features that differentiate the categories, and participants seemed to be mindful of this advantage. Furthermore, an increased degree of within-similarity switching was associated with better performance in a later categorization test.

Table 7 Measures of interleaving across the four quarters

Measure	Experiment quarter			
	1	2	3	4
Proportion of switching	.53	.44	.46	.50
Similar switch ratio	.47	.53	.58	.45

These findings show that learners are strategic in their study choices and do not always stick to a predominantly blocked strategy. While our participants still showed an inclination toward blocked study (as in Tauber et al., 2013), our results paint a more nuanced picture of their strategy rather than an overwhelming preference for one or the other (see also Kornell & Vaughn, 2018). Our results are novel in showing that the preference toward blocking or interleaving can be modified by category similarity. Also, self-regulated learners performed better than learners in experimenter-controlled study conditions (whether blocked or interleaved). Furthermore, participants' choice behaviors during study predicted subsequent performance, which is consistent with the notion that learning is influenced by one's study behaviors and that there are ideal study behaviors which optimize learning.

Choosing to block versus interleave during study

Prior research has shown a strong preference for blocked study (Kornell & Bjork, 2008; Yan et al., 2016; Yan, Soderstrom, Seneviratna, Bjork, & Bjork, 2017). Our paper does not refute earlier claims that learners prefer blocking; indeed, across both experiments, the proportion of "blocking" chosen by participants on every trial was greater than predicted by random chance. What we have shown is that participants do choose to interleave when it is helpful (i.e., when the to-be-learned categories are highly similar and hard to discriminate). We have strong evidence that this behavior is strategic; learners tend to interleave selectively among similar (and not dissimilar) categories (Experiment 2).

Our results extend previous research in a number of important ways. First, our paradigm allowed participants to make choices during real-time study; in most previous studies, participants were asked to rate the effectiveness of a particular study schedule, or were asked to choose between hypothetical scenarios (e.g., Kornell & Bjork, 2008; McCabe, 2011; Yan et al., 2016; Yan et al., 2017). Second, we showed that the choices learners make during study can predict their performance in a subsequent test. These results complement the results of earlier studies indicating that although participants do show an inclination for blocking over interleaving, they are capable of shifting toward a more interleaved strategy when it is beneficial. As between-category similarity may have been much more salient to learners studying our stimuli (both the blobs and rocks), it is perhaps unsurprising that they tended to interleave more than the learners in the Tauber et al. (2013) experiments. We also required participants to learn fewer categories (four or six) in our experiments, which may have made interleaving between categories a less challenging prospect than in previous studies (e.g., the eight or 12 in Tauber et al., 2013, or 10 in Kornell & Vaughn, 2018). If there are many more categories to learn, participants may be more inclined to block those categories because that choice may be easier to make and keep track of.

It should be noted that we are not claiming that learners always make ideal study choices when learning categories (in fact, even in our own data, there was substantial variability). Our results do show, however, that learners' strategic choices are influenced by the structure of the categories they are confronted with (when it is highly salient, as in our case). In fact, our data show support for both learners having a preference for blocked study, while also being capable of making strategic choices that benefit learning. In the metamemory literature, there are many instances of inaccurate monitoring of learning or illusions of competence (see Bjork, Dunlosky, & Kornell, 2013, for a review). Yet there is also ample evidence of effective monitoring and control resulting in better memory (Kornell & Metcalfe, 2006; Son, 2004, 2010). Indeed, the Kornell and Vaughn (2018) results indicate that learners can be quite sophisticated when making study choices, showing a desire to be "fair" to each category when sampling them for study.

Ultimately, metacognitive judgments are inferential in nature and based on various beliefs and cues; the accuracy of the judgments (or decisions) depend on which factors are salient during the learning context and whether the factors relied upon by the learner are predictive of learning outcomes (Koriat, 1997).

The benefits of self-regulation in learning

The results of Experiment 1 showed that learners who were allowed choice in the study task performed better than both the interleaved and blocked control participants. In their review, Gureckis and Markant (2012) discuss some possible explanations for the benefit of self-regulated learning over passive or yoked controls. On one hand, self-regulated participants may have "data driven" advantages (i.e., learners can optimize their learning experience by reducing uncertainty and avoiding redundant information through their choices). For example, Markant and Gureckis (2010) found that self-directed participants tended to avoid redundant exemplars that they could already confidently classify in a novel perceptual category learning task, and Tullis and Benjamin (2011) showed that self-paced participants who allocate their time in a manner consistent with a discrepancy-reduction strategy performed better than controls in a memory task. In addition, self-regulated participants may also enjoy "decision driven" advantages, which relate to the psychological benefits of choice and active exploration, as well as the potential for enhanced task engagement (Gureckis & Markant, 2012; Leotti, Iyengar, & Oshsner, 2010).

The results of Experiment 2 extend the idea of a "data driven" advantage by showing that learners who make more optimal choices during study perform better in a subsequent test. In our study, self-regulated participants may reduce uncertainty by choosing to study the same category over and over if they wish to verify the presence of a defining within-category feature, or by choosing a different category in order to juxtapose different between-category features, depending on the hypotheses that they currently hold. As for the "decision-driven" advantages, research in the social psychology literature has repeatedly demonstrated that giving participants choice and autonomy has a positive impact through increasing intrinsic motivation, perceived control, and task performance (Cordova & Lepper, 1996; Deci & Ryan, 1985; Dember, Galinsky, & Warm, 1992). Even when the "choices" afforded to participants are illusory (Dember et al., 1992) or minimal and pedagogically irrelevant (Cordova & Lepper, 1996), allowing them any choice at all increases performance in the task, suggesting that the mere perception of choice increases task motivation.

Future directions

Future work should attempt to disentangle the "data-driven" from "decision-driven" advantages of choice in category learning. Some potential avenues are using the honor/dishonor paradigm (Kornell & Metcalfe, 2006), where participants' choices are either honored or dishonored, or running yoked control conditions, where one participant is allowed to choose, but a second participant also must follow those choices.

Moreover, while the current investigation focused on discriminative contrast (Carvalho & Goldstone, 2014; Kornell & Bjork, 2008), there is evidence that temporal spacing of categories (Cepeda, Pashler, Vul, Wixted, & Rohrer, 2006) also plays a role in the interleaving benefit, at least for learning mathematics problems (Foster, Mueller, Was, Rawson, & Dunlosky, 2019). Although our similarity manipulations were focused on interleaving as a way of promoting discriminative contrast, rather than spacing, our results do not preclude the possibility that the discriminative contrast provided by interleaving might play a larger role in learning the easily confused perceptual categories such as those in the current work, while spacing may play a larger role in other tasks, such as mathematics problem solving. Future work should explore the potential benefits of both discriminative contrast and spacing on various tasks in the context of self-regulated learning.

Finally, although our choice behavior analysis was exploratory, there were indications that participants shifted their choice behavior over the course of learning. Our results showed that participants tended to interleave more at the beginning of the study phase (Experiments 1 and 2), while they tended to interleave more among similar categories in the middle of the study phase (Experiment 2). This behavior would be consistent with adapting an initial exploration strategy in which the structure of the categories is discovered, and a subsequent strategy of optimizing category switching based on the earlier gathered information. These results provide some preliminary support for the Kornell and Vaughn (2018) hypothesis that learners are "foraging" for information from different categories when self-regulating their learning. On the other hand, when Yan et al. (2017) asked participants to choose hypothetical study sequences, they tended to prefer a sequence that started out with blocked presentation, but shifted toward interleaving over time, suggesting that hypothetical preferences may not align with actual behavior during learning. Future work should explore the factors that influence strategy shifting over time, as well as the extent to which this behavior is strategic or predicts performance.

Conclusions

Learners can be strategic during category learning, choosing to interleave among categories more often when learning categories that are highly similar (as opposed to blocking their study by category), and such study behavior predicts better learning outcomes. We also found that learners who were allowed to choose the sequencing of categories during study outperformed learners whose study sequencing was experimenter controlled. It is important for future research to explore further how sequencing effects on category learning might interact with not only the type of learning materials, but also learner choice and agency.

Acknowledgements We acknowledge Amanda Chen, Autumn Diggs, Danielle Fang, Nandita Kasireddy, and Yiu Chung Wong for their assistance in data collection.

Open practices statement All stimuli, data, and analysis files for both experiments are available (https://osf.io/6mrg2/).

References

Ariel, R., & Dunlosky, J. (2013). When do learners shift from habitual to agenda-based processes when selecting items for study? *Memory & Cognition, 41*(3), 416–428.

Ashby, F. G., & Maddox, W. T. (2011). Human category learning 2.0. *Annals of the New York Academy of Sciences, 1224,* 147–161.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48.

Bjork, R. A., Dunlosky, J., & Kornell, N. (2013). Self-regulated learning: Beliefs, techniques, and illusions. *Annual Review of Psychology*, *64*, 417–444.

Brunmair, M., & Richter, T. (2019). Similarity matters: A meta-analysis of interleaved learning and its moderators. *Psychological Bulletin, 145*(11), 1029–1052.

Brysbaert, M., & Stevens, M. (2018). Power analysis and effect size in mixed effects models: A tutorial. *Journal of Cognition*, *1*(1), 9. doi: https://doi.org/10.5334/joc.10

Carvalho, P. F., & Goldstone, R. L. (2014). Putting category learning in order: Category structure and temporal arrangement affect the benefit of interleaved over blocked study. *Memory & Cognition, 42*(3), 481–495.

Cepeda, N. J., Pashler, H., Vul, E., Wixted, J. T., & Rohrer, D. (2006). Distributed practice in verbal recall tasks: A review and quantitative synthesis. *Psychological Bulletin, 132*(3), 354–380.

Cordova, D. I., & Lepper, M. R. (1996). Intrinsic motivation and the process of learning: Beneficial effects of contextualization, personalization, and choice. *Journal of Educational Psychology*, *88*(4), 715–730.

Deci, E. L., & Ryan, R. M. (1985). *Intrinsic motivation and self-determination in human behavior*. New York, NY: Plenum.

Dember, W. N., Galinsky, T. L., & Warm, J. S. (1992). The role of choice in vigilance performance. *Bulletin of the Psychonomic Society*, *30*(3), 201–204.

Foster, N. L., Mueller, M. L., Was, C., Rawson, K. A., & Dunlosky, J. (2019). Why does interleaving improve math learning? The contributions of discriminative contrast and distributed practice. *Memory & Cognition*, *47*(6), 1088–1101.

Green, P., & MacLeod, C. J. (2016). SIMR: An R package for power analysis of generalized linear mixed models by simulation. *Methods in Ecology and Evolution*, *7*(4), 493–498.

Gureckis, T. M., & Markant, D. B. (2012). Self-directed learning: A cognitive and computational perspective. *Perspectives on Psychological Science*, *7*(5), 464–481.

Johnson, P. C., Barry, S. J., Ferguson, H. M., & Müller, P. (2015). Power analysis for generalized linear mixed models in ecology and evolution. *Methods in Ecology and Evolution,* 6(2), 133–142.

Kang, S. H. K. (2017). The benefits of interleaved practice for learning. In J. C. Horvath, J. Lodge, & J. A. C. Hattie (Eds.), *From the laboratory to the classroom: Translating science of learning for teachers* (pp. 79–93). New York, NY: Routledge.

Kang, S. H. K., & Pashler, H. (2012). Learning painting styles: Spacing is advantageous when it promotes discriminative contrast. *Applied Cognitive Psychology*, *26*(1), 97–103.

Koriat, A. (1997). Monitoring one's own knowledge during study: A cue-utilization approach to judgments of learning. *Journal of Experimental Psychology: General*, *126*(4), 349–370.

Kornell, N., & Bjork, R. A. (2008). Learning concepts and categories is spacing the "enemy of induction"?. *Psychological Science, 19*(6), 585–592.

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

Kornell, N., & Metcalfe, J. (2006). Study efficacy and the region of proximal learning framework. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 32*(3), 609–622.

Kornell, N., & Vaughn, K. E. (2018). In inductive category learning, people simultaneously block and space their studying using a strategy of being thorough and fair. *Archives of Scientific Psychology*, *6*(1), 138–147.

Leotti, L. A., Iyengar, S. S., & Ochsner, K. N. (2010). Born to choose: The origins and value of the need for control. *Trends in Cognitive Sciences*, *14*(10), 457–463.

Markant, D., & Gureckis, T. (2010). Category learning through active sampling. In S. Ohlsson & R. Catrambone (Eds.), *Proceedings of the Annual Meeting of the Cognitive Science Society* (pp. 248–253). Austin, TX: Cognitive Science Society.

Maynard, L. (2006). The role of repetition in the practice sessions of artist-teachers and their students. *Bulletin of the Council for Research in Music Education, 167,* 61–72.

McCabe, J. (2011). Metacognitive awareness of learning strategies in undergraduates. *Memory & Cognition*, *39*(3), 462–476.

Morehead, K., Rhodes, M. G., & DeLozier, S. (2016). Instructor and student knowledge of study strategies. *Memory*, *24*(2), 257–271.

Nosofsky, R. M., Sanders, C. A., Meagher, B. J., & Douglas, B. J. (2017). Toward the development of a feature-space representation for a complex natural category domain. *Behavior Research Methods, 50*(2), 530–556.

Rohrer, D., Dedrick, R. F., & Hartwig, M. K. (in press). The scarcity of interleaved practice in mathematics textbooks. *Educational Psychology Review*.

Rohrer, D., Dedrick, R. F., Hartwig, M. K., & Cheung, C.-N. (2019). A randomized controlled trial of interleaved mathematics practice. *Journal of Educational Psychology, 112*(1), 40–52.

Rohrer, D., Dedrick, R. F., & Stershic, S. (2015). Interleaved practice improves mathematics learning. *Journal of Educational Psychology, 107*(3), 900–908.

Sana, F., Yan, V. X., & Kim, J. A. (2017). Study sequence matters for the inductive learning of cognitive concepts. *Journal of Educational Psychology*, *109*(1), 84–98.

Son, L. K. (2004). Spacing one's study: evidence for a metacognitive control strategy. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *30*(3), 601–604.

Son, L. K. (2010). Metacognitive control and the spacing effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *36*(1), 255–262.

Tauber, S. K., Dunlosky, J., Rawson, R. A., Wahlheim, C. N., & Jacoby, L. L. (2013). Self-regulated learning of a natural category: Do people interleave or block exemplars during study? *Psychonomic Bulletin & Review*, 20, 356–363.

Tullis, J. G., & Benjamin, A. S. (2011). On the effectiveness of self-paced learning. *Journal of Memory and Language*, *64*(2), 109–118.

Verkoeijen, P., & Bouwmeester, S. (2014). Is spacing really the "friend of induction"? *Frontiers in Psychology*, *5*, 259.

Wahlheim, C. N., Dunlosky, J., & Jacoby, L. L. (2011). Spacing enhances the learning of natural concepts: An investigation of mechanisms, metacognition, and aging. *Memory & Cognition, 39*(5), 750–763.

Yan, V. X., Bjork, E. L., & Bjork, R. A. (2016). On the difficulty of mending metacognitive illusions: A priori theories, fluency effects, and misattributions of the interleaving benefit. *Journal of Experimental Psychology: General*, *145*(7), 918–933.

Yan, V. X., Soderstrom, N. C., Seneviratna, G. S., Bjork, E. L., & Bjork, R. A. (2017). How should exemplars be sequenced in inductive learning? Empirical evidence versus learners' opinions. *Journal of Experimental Psychology: Applied*, *23*(4), 403–416.