Instructional Intervention Effects on Interleaving Preference and Distance During Self-Regulated Inductive Learning

Anran Li[1], Mengqi Hu[1], Aohan Xu[2], Wenbo Zhao[3], Xiao Hu[2], David R. Shanks[4], Liang Luo[1,5], Chunliang Yang[1,6]

[1] Institute of Developmental Psychology, Faculty of Psychology, Beijing Normal University, China.

[2] Faculty of Psychology, Beijing Normal University, China.

[3] School of Social Development and Public Policy, Beijing Normal University, Beijing, China.

[4] Division of Psychology and Language Sciences, University College London, London, the UK.

[5] State Key Laboratory of Cognitive Neuroscience and Learning, Beijing Normal University, Beijing, China.

[6] Beijing Key Laboratory of Applied Experimental Psychology, National Demonstration Center for Experimental Psychology Education, Beijing Normal University, China.

Author Note

Correspondence concerning this article should be addressed to Chunliang Yang (chunliang.yang@bnu.edu.cn), Institute of Developmental Psychology, Faculty of Psychology, Beijing Normal University, 19 Xinjiekouwai Street, Haidian District, Beijing 100875, China.

All data and experimental stimuli have been made publicly available via the Open Science Framework (OSF) at https://osf.io/9w36k/.

Acknowledgement

This research was supported by the Natural Science Foundation of China (32000742; 32171045; 32200841), the Research Program Funds of the Collaborative Innovation Center of Assessment toward Basic Education Quality at Beijing Normal University (2021-01-132-BZK01), the UK Economic and Social Research Council (ES/S014616/1), and the Fundamental Research Funds for the Central Universities (2022NTSS36).

Abstract

Interleaving (intermixing exemplars from different categories) is more effective in promoting inductive learning than blocking (massing exemplars from a given category together). Yet learners typically prefer blocking over interleaving during self-regulated inductive learning, highlighting the need to develop effective interventions to overcome this metacognitive illusion and promote learners' practical use of the interleaving strategy. Drawing on a sample of university students, three experiments examined the effects of an instructional intervention on (1) correction of metacognitive fallacies regarding the superiority of blocking over interleaving for inductive learning, (2) adoption of the interleaving strategy during self-regulated learning when learners are allowed to make study choices exemplar-by-exemplar, (3) classification performance, and (4) transfer of category learning across different domains. Experiments 1 and 2 showed that instructions about the benefits of interleaving over blocking improved metacognitive awareness of the efficacy of interleaving and enhanced self-usage of the interleaving strategy during learning of new categories. However, this intervention had negligible influence on interleaving distance and did not improve classification performance. Experiment 3 found that informing learners about the benefits of extensive interleaving, as compared to minimal interleaving or no interleaving, successfully increased interleaving distance and boosted classification performance, and the intervention effects transferred to learning categories in a different domain. These findings support the practical use of the instructional intervention in promoting self-usage of the interleaving strategy, and highlight the important role of enlarging interleaving distance in facilitating inductive learning.

Educational Impact and Implications Statement

Educators have endeavored to explore methods to enhance the efficiency of inductive learning. Intermixing exemplars from different categories produces superior learning outcomes than studying each

category's exemplars in a blocked manner. However, most learners fail to appreciate the benefits of interleaving and prefer to studying exemplars in a massed schedule. Providing learners with individualized feedback on blocked vs. interleaved performance and highlighting relevant findings from previous research can effectively enhance their awareness of the merits of interleaving and encourage them to adopt the interleaved strategy during self-regulated learning. When learners attempt to intersperse exemplars of a given category with those from many other categories, their inductive performance improves.

Keywords: Inductive learning; Interleaving; Metacognitive illusion; Instructional intervention; Self-regulated learning

Inductive learning refers to generalizing characteristic features from a series of exemplars so that the world can be simplified and understood (Holland et al., 1989). Different strategies have been developed to promote the efficiency of inductive learning, such as testing (Jacoby et al., 2010; Yang & Shanks, 2018), fading (Pashler & Mozer, 2013) and feature highlighting (Miyatsu et al., 2019). Besides the strategies listed, interleaving – intermixing exemplars from different categories – has been identified as another powerful tool to foster inductive learning (for meta-analyses, see Brunmair & Richter, 2019; Firth et al., 2021).

As first reported by Kornell and Bjork (2008), interleaving exemplars from different categories (e.g., A_1 B_1 C_1 A_2 B_2 C_2 A_3 B_3 C_3; letters represent different categories, with subscripts denoting different exemplars in a given category) tends to produce superior inductive learning outcomes by comparison with blocking exemplars by category (e.g., A_1 A_2 A_3 B_1 B_2 B_3 C_1 C_2 C_3). In their Experiment 1a, participants were instructed to study 6 paintings by each of 12 artists in order to master their painting styles. Half of the artists' paintings were studied under a massed condition in which all 6 paintings from a given artist were presented in succession within a block. In contrast, one painting from each of the remaining 6 artists was intermixed within a block in an interleaved condition. After the study phase, all participants undertook an induction test in which 48 new paintings (i.e., 4 from each artist) were shown one-by-one, and participants judged which artist painted each one. Kornell and Bjork (2008) found that the mean rate of correct classifications was significantly higher in the interleaved than in the blocked condition, reflecting an *interleaving effect* on inductive learning.

The interleaving effect has been successfully replicated by dozens of studies which have further extended it to learning of animal species (Birnbaum et al., 2013; Kornell & Vaughn, 2018; Wahlheim et al., 2011), chemical compounds (Eglington & Kang, 2017), mathematical volume calculations (Foster et

al., 2019; Rohrer & Taylor, 2007), cognitive and social concepts (Rawson et al., 2015; Sana et al., 2017), musical styles and intervals (Wong et al., 2021; Wong et al., 2020), second language syntax (Nakata & Suzuki, 2019; Suzuki et al., 2020) and other complex materials in educational settings (Mielicki & Wiley, 2022). Interleaving-enhanced inductive learning also persists across a span of study-test intervals ranging from seconds (Kornell & Bjork, 2008; Kornell et al., 2010; Verkoeijen & Bouwmeester, 2014; Wahlheim et al., 2011), days (Pan et al., 2019; Taylor & Rohrer, 2010) and weeks (Zulkiply, 2013; Zulkiply & Burt, 2013b), to months (Rohrer et al., 2015). Accordingly, interleaving is a recommended strategy to organize exemplars with respect to inductive learning (for a recent systematic review, see Firth et al., 2021).

Extending interleaving distance to optimize the benefits of interleaving

To maximize the benefits of interleaved study, theoretical accounts for the interleaving effect should be leveraged. A well-established theory is the *discriminative-contrast hypothesis* postulating that interleaving facilitates category learning through promoting comparison and contrast of differences between categories (search-for-difference) (Kang & Pashler, 2012). This hypothesis predicts that the degree of similarity between categories is a critical factor that moderates the magnitude of the interleaving effect (Brunmair & Richter, 2019). Indeed, previous studies confirmed that the more similar the categories are, the more efficient interleaving is for enhancing category learning (Carvalho & Goldstone, 2014a, 2014b; Zulkiply & Burt, 2013a). However, there is evidence showing that the discriminative-contrast hypothesis itself cannot completely explain the interleaving effect.

To illustrate, Yan and Sana (2019) found that the interleaving effect was not abolished even when the differences between categories were sufficiently salient to notice. Specifically, in the study phase of their Experiment 1, participants viewed artworks from four artists, with each artist's set of eight paintings consistently focusing on the same object, with each object differing from those portrayed by all other

artists. That is, Artist 1's paintings were always of buildings, Artist 2's paintings were always of flowers, and so on. As such, paintings from different artists were readily distinguishable. Yan and Sana's findings thus suggest that in addition to juxtaposing different categories, other properties inherent in interleaved sequences may also contribute to the interleaving effect. Similarly, the interleaving effect survived when Rohrer et al. (2014) directed middle school students to practice four kinds of superficially dissimilar math problems in both interleaved and blocked sequences. Moreover, Eglington and Kang (2017) eliminated the necessity for between-category comparison entirely by highlighting the diagnostic features of each category, yet the interleaved schedule still produced superior induction performance to the blocked one. These findings jointly imply that discriminative contrast is not the only mechanism underlying the interleaving effect.

An alternative but not mutually exclusive explanation of the interleaving effect is based on distributed retrieval from long-term memory (for a review, see Cepeda et al., 2006). According to this *study-phase retrieval hypothesis*, repeated exposure to a category triggers retrieval of its previous instances and this act of retrieval is the key to promoting learning (Shanks et al., 2023; Yang et al., 2021). Specifically, longer temporal intervals between same-category exemplars afforded by interleaved schedules induce more forgetting and thereby more effortful retrieval from memory. In a blocked schedule, in contrast, previous exemplars are highly accessible in working memory and hence the retrieval process is not demanding when a new exemplar is encountered (Dunlosky et al., 2013).

As effective strategies typically entail high mental demands – the so-called "desirable difficulties" (Bjork, 1994) framework – increasing retrieval difficulty by extending temporal distance between category repetitions in the interleaved schedule is expected to further enhance the efficacy of interleaving. This prediction has been confirmed by Birnbaum et al.'s (2013) Experiment 3. In this experiment, two

(small vs. large interleaving) groups of participants studied 16 species of butterflies. For the small interleaving group, exemplars were interleaved across only 4 species. By contrast, for the large interleaving group, exemplars were interleaved across all 16 species. Birnbaum et al. (2013) observed that classification performance was substantially better in the large than in the small interleaving group, lending support to the study-phase retrieval hypothesis.

It is, however, crucial to note that large interleaving is not only characterized by large temporal distance, but also allows participants to search for differences among a larger number of categories, thus improving discriminative-contrast and facilitating category learning. Therefore, besides inducing more effortful retrieval, large interleaving also facilitates a more thorough and comprehensive differentiation. In light of this, the superiority of large interleaving (relative to small interleaving) may not only result from study-phase retrieval but also derive from enhanced discriminative contrast.

From the arguments discussed above, it is evident that maximizing the benefits of interleaving through intermixing exemplars from a large number of categories should fully leverage both the discriminative-contrast and study-phase retrieval mechanisms. Yet triggering demanding retrieval by simply interspersing filler tasks between exemplars in an interleaved schedule may prove counterproductive, as it disrupts the discrimination process (Sana et al., 2017; Zulkiply & Burt, 2013a). Inspired by Birnbaum et al. (2013), separating exemplars from a given category with those from other categories could be an approach to avoid interrupting discrimination and meanwhile induce more effortful retrieval. Hence, we introduce *interleaving distance* (i.e., the mean number of intervening exemplars between exemplars from a given category) as a new indicator to measure the degree of interleaving in self-regulated learning. It is important to note that the definition of "interleaving distance" in the current study is similar to the "exemplar-to-exemplar spacing" metric used in Kornell and Vaughn's (2018)

Experiment 2, yet subtle differences exist in the calculation approaches. We avoid the term "spacing" for this metric because, in addition to capturing the temporal intervals between exemplars from the same category, "interleaving distance" also emphasizes the spatial juxtaposition between different categories.

Metacognitive illusions and underemployment of interleaving

Besides the interleaving effect per se, people's metacognitive awareness about the benefits of interleaving has also attracted substantial research interest. Intriguingly, myriad studies consistently showed that, although actual performance exhibited strong interleaving effects in inductive learning, a majority of participants erroneously believed that blocked learning is more effective, reflecting that learners typically lack metacognitive awareness about the benefits of interleaving (Birnbaum et al., 2013; Gluckman et al., 2014; Kornell & Bjork, 2008; Kornell et al., 2010; Yan et al., 2016; Yan et al., 2017; Zulkiply et al., 2012). For instance, at the end of the induction test in Kornell and Bjork's (2008) Experiment 1a, participants were asked to report which schedule they thought helped them learn better by selecting one from three options: "massed" (i.e., blocked), "about the same", and "spaced" (i.e., interleaved). Even though 78% of participants showed superior classification performance in the interleaved than in the blocked condition, 63% of them erroneously judged that blocking was superior to interleaving. Even among older adults with rich experience of learning, 75% perceived blocking as more effective than interleaving (Kornell et al., 2010). This metacognitive fallacy associated with interleaving has been attributed to a range of mechanisms including processing fluency, prior beliefs, and learners' wishful thinking about their uniqueness (for detailed discussion, see Yan et al., 2016).

It is well-known that individuals' metacognitive judgements about the relative efficacy of learning strategies (i.e., metacognitive monitoring) guide metacognitive control (Yang et al., 2017; Yang et al., 2018), such as study time allocation (e.g., Metcalfe & Kornell, 2003) and study strategy selection (e.g.,

Karpicke, 2009). Thus, metacognitive illusions associated with the interleaving effect could result in underemployment of the interleaving strategy during self-regulated learning.

Tauber et al. (2013) provided a clear illustration of exactly this in a self-regulated learning task which closely simulated a real-life learning situation with high ecological validity. Participants were shown the names of eight bird families and instructed to choose which category they would like to study next. Then an exemplar from the chosen category was randomly selected and presented on screen for studying. To quantify strategy use, Tauber et al. (2013) estimated how often participants blocked their study of exemplars from the same bird family. The amount of blocking was calculated as the frequency that two exemplars from the same category were studied consecutively. They observed that participants blocked a majority of category exemplars (ranging from 70% to 97% in Experiments 1-4) during self-regulated learning. Tauber et al. (2013) also calculated the proportion of "blockers", defined as the percentage of participants who scheduled over 50% of exemplars in a blocked sequence, and found that most participants (ranging from 78% to 100% in Experiments 1-4) were classified as blockers. Jointly, these findings manifest extreme underemployment of interleaving during self-regulated learning. Using similar paradigms and study sequence metrics, Kornell and Vaughn (2018) observed that in a self-regulated learning task in which participants studied different species of penguins, they blocked only 47.2% of the exemplars. Nevertheless, this blocking rate still significantly exceeded the chance level of 15.6%.

Despite learners' prevalent preference for high proportions of blocking, evidence suggests that in some situations they spontaneously interleave their study sequences for discrimination purposes, especially when the need for between-category comparisons is made sufficiently salient. For instance, when motivated to minimize confusion, learners strategically alternated between two superordinate

mushroom categories to a higher extent (Abel, 2023). In a similar vein, with category similarity manipulated within-subjects, Lu et al. (2021) found a tendency among learners to interleave between highly similar rock categories.

Even though learners are capable of incorporating interleaving in their study sequence choices when the importance of detecting differences among categories is highlighted, their preference for blocking exemplars continues to be deeply ingrained. For instance, in Able (2023), participants' blocking rate actually reached as high as 80%, even when they were faced with a survival threat to distinguish between edible and poisonous mushrooms (i.e., when a high motivation to avoid confusion is induced). The entrenched habit to block was also observed in a study by Lu et al. (2021) in which participants had three options before studying each subsequent exemplar: proceeding with the same category, switching to a similar category, or switching to a dissimilar category. Although switching to a similar category would presumably be most beneficial, the proportion of blocked choices still exceeded 50%. In brief, learners' underemployment of interleaving is pervasive across diverse circumstances, irrespective of whether searching for differences among categories is emphasized as a key learning goal.

Overall, metacognitive illusions associated with interleaving lead to excessive use of ineffective strategies which exposes learners to blocked schedules, in turn producing poor inductive learning efficiency. Hence, it is important to correct learners' erroneous beliefs in order to improve self-usage of interleaving and facilitate category learning.

Interventions for correcting metacognitive illusions and improving interleaving strategy usage

Although metacognitive illusions associated with interleaving are prevalent among learners, to our knowledge, only three studies have explored how to alleviate or reverse these illusions (Onan et al., 2022; Sun et al., 2022; Yan et al., 2016). Yan et al. (2016) tried to correct participants' misaligned metacognitive

knowledge by drawing their attention to the connection between the study schedule (blocked vs. interleaved) and their own performance. Disappointingly, their Experiment 5 showed that, even when the experimenters (1) prompted participants to establish an unambiguous connection between paintings and their presentation schedules by surrounding them within distinct frames for each schedule, (2) provided corrective feedback about the artist responsible for each new painting during the test phase, (3) offered explicit instructions informing participants that interleaving is a more effective strategy than blocking for most (i.e., 90%) people, and (4) explained the reasons why interleaving is more powerful than massing, only 36% of the participants reported that interleaving is better than blocking after the intervention.

The second intervention study was conducted by Sun et al. (2022), who developed another intervention to promote metacognitive awareness of the interleaving effect, increase self-use of the interleaving strategy, and improve classification performance. Their experiments were composed of two learning tasks, with an intervention phase interpolated between them. The pre-intervention task was identical to that in Kornell and Bjork's (2008) Experiment 1a, in which two (intervention vs. control) groups of participants studied 6 artists' paintings in a blocked manner, with another 6 artists' paintings studied in an interleaved manner. Then, both groups took an induction test. After the test, they reported which strategy they thought was more effective. Next, participants in the intervention group were shown their own classification performance under interleaved and blocked conditions as well as a summary of the findings from Kornell and Bjork (2008) as a briefing designed to convince them that interleaving is more effective than blocking. By contrast, participants in the control group read comparable instructions irrelevant to the interleaving effect.

After the intervention phase, both groups again reported which strategy they thought was more effective for inductive learning and then performed a post-intervention task in which they were allowed to

freely choose which strategy (interleaving vs. blocking) they would like to employ to study each new category (i.e., new artists' painting styles or butterfly species). The names of the to-be-studied categories were displayed one at a time, and participants decided whether they wanted to study a given category in an interleaved or blocked manner. Their choices were then honored in the following study phase and the procedure in the test phase was the same as that in the pre-intervention task.

Sun et al. (2022) found that their instruction intervention was successful in (1) enhancing participants' appreciation of the benefits of interleaving, (2) motivating participants to employ the interleaving strategy more frequently at the category level in the post-intervention task (see below for detailed discussion), and (3) improving classification accuracy in the post-intervention task. Moreover, Sun et al. (2022) showed that the intervention effect on interleaving strategy usage was somewhat long-lasting (i.e., at least 24 hours) and transferable from learning painting styles to learning butterfly species.

Aside from the instructional intervention, Onan et al. (2022) developed another approach to promote the use of the interleaving strategy with an eye towards subjective experiences. Onan et al. (2022) argued that students' inaccurate monitoring of on-task experiences contributed to their poor strategy choices. Repeated rating of invested effort and perceived learning during study helped students keep track of how their subjective experience changed across time, so they could recognize that their perceived learning increased as their perceived effort decreased for interleaved practice. As a result, students partially zoomed out of the immediate learning gains afforded by blocking and were prompted to choose effortful yet effective interleaving (for related findings, see Janssen et al., 2023).

Rationale and overview of the current study

Although previous intervention studies have attained preliminary success in improving students' metacognitive appreciation about the efficacy of interleaving and enhanced self-employment of

interleaving strategy, several important issues still need to be addressed.

First, the learning tasks employed in previous intervention studies lack ecological validity and are not representative of real-life learning settings in which learners can freely make strategy choices after studying each exemplar. For instance, in the post-intervention task of Sun et al. (2022), participants were asked to make a binary choice for each category regarding whether they would like to study all of its exemplars in a blocked or an interleaved manner (i.e., making strategy choices at the category level). Once these category-level choices had been made, they were forced to study each category's exemplars in the pre-chosen manner and could not make any strategy changes during the study phase. That is, participants had to study all exemplars from the same category in the same manner according to their category-level strategy choices made before the study phase. Likewise, Onan et al. (2022) asked participants to report their endorsed strategy to study novel materials (i.e., making a one-off decision about either interleaving or blocking) and all to-be-learned categories were scheduled either in pure blocking or interleaving according to their choice. Consequently, participants passively experienced exemplars displayed on the screen, unable to determine the category they wanted to study next.

The learning tasks employed by Sun et al. (2022) and Onan et al. (2022) are not representative of everyday learning scenarios, in which learners typically make a strategy choice after studying each exemplar (i.e., making a choice regarding whether they want next to study a new exemplar from the same or a different category), as in the study by Tauber et al. (2013). Because of this, the practical implications of previous intervention findings are highly limited to experimenter-controlled learning, but not generalizable to self-regulated learning. Furthermore, it has to be highlighted that strategy choices made at the category level or on a one-off basis may not be representative of the actual use of that strategy. Previous studies clouded the construct of strategy choice and strategy implementation. Sun et al. (2022)

supposed that in the post-intervention learning task participants would regulate their study schedule strictly following their category-level choices (or study plans) made prior to the study phase. Unfortunately, in many situations, learners do not always implement their study plans. For instance, Badali et al. (2022) found that, even though participants made a clear plan to recall difficult items more times before dropping them from a study list, actually the number of correct recalls during the study phase was roughly equal between easy and difficult items, suggesting that learners do not always commit to their study plans during self-regulated learning.

Executing a plan may prove to be more challenging when to-be-studied items are presented in a sequential format—the very format used in the exemplar-by-exemplar self-regulated learning task—compared to a simultaneous format (Ariel et al., 2009; Dunlosky & Thiede, 2004; Middlebrooks & Castel, 2018). In fact, learners' regulation of study behaviors relies more on their perceived knowledge state during the study phase rather than study plans (Koriat & Bjork, 2005; Koriat & Bjork, 2006). As such, the findings documented by Sun et al. (2022) and Onan et al. (2022) cannot be used to directly guide self-regulated learning practices.

Finally, but importantly, merely enhancing interleaving frequency (i.e., interleaving usage) in a self-regulated learning task may be insufficient to boost learning performance. As discussed above, the magnitude of the interleaving effect is dependent on interleaving distance: large interleaving is more beneficial than small interleaving. Critically, frequent interleaving does not promise large interleaving distance. For example, sequence ABABCDCD (switch frequency = 100%) has a larger interleaving frequency than sequence ABCDDCBA (switch frequency = 86%), but the former is associated with a smaller interleaving distance (distance = 1) than the latter (distance = 3; see below for details of how to calculate interleaving distance). Given that interleaving distance cannot be calculated in Sun and

colleagues' (2022) study as their post-intervention task was not truly self-regulated, it remains unclear whether improving metacognitive awareness of the benefits of interleaving can also increase interleaving distance. More importantly, how to improve interleaving distance in a self-regulated inductive learning task is yet to be explored.

Overall, previous intervention studies lack ecological validity and their documented findings cannot be used to guide educational practices. It remains unknown whether the instruction intervention developed by Sun et al. (2022) can promote self-usage of the interleaving strategy and enhance inductive learning outcomes when learners are allowed to make choices of blocking versus interleaving at the exemplar level. Moreover, it has yet to be explored whether the instruction intervention about the interleaving effect can improve interleaving distance during self-regulated learning, which is a critical factor determining the magnitude of interleaving-enhanced inductive learning.

The current study aims to address these gaps. Specifically, Experiment 1 employed the same procedure and instructional intervention as those in Sun et al.'s (2022) Experiment 1, but introduced a critical difference. That is, in the post-intervention task, the intervention and control groups did not make category-level strategy choices (interleaving vs. blocking). Instead, they were instructed to freely choose on each trial which artist's painting they wanted to study next. Interleaving frequency, percentage of interleavers, and interleaving distance were calculated as measures of self-usage of the interleaving strategy. Experiment 2 sought to determine whether the intervention effect on self-usage of interleaving can transfer to learning of categories in other domains. To that end, participants were asked to study categories in another domain (i.e., butterfly species) in the post-intervention task. Going beyond Experiments 1 and 2, Experiment 3 explored whether highlighting the benefits of large over small interleaving in the instructions can enhance interleaving distance and improve inductive learning

outcomes.

Experiment 1

Experiment 1 investigated whether the instruction intervention developed by Sun et al. (2022), in which individually-tailored feedback on blocked vs. interleaved score is provided together with a summary of research findings about the interleaving effect from prior studies, can (1) correct metacognitive bias, (2) enhance adoption of interleaving during self-regulated learning, (3) increase interleaving distance, and (4) facilitate category learning performance.

Method

Participants

Sun et al. (2022) observed Cohen's $d = 0.96$ for the intervention effect on participants' category-level strategy choices in their Experiment 1. Accordingly, a power analysis was conducted via G*Power 3 (Faul et al., 2007) for an independent-samples t test ($\alpha = .05$, 2-tailed). This determined that at least 24 participants in each group were required to detect an intervention effect on strategy choices at a power of .90. To ensure sufficient statistical power and following Sun et al. (2022), we pre-planned to recruit 30 participants in each group.

Finally, a total of 63 participants (M age = 21.98, SD = 1.98; 70% female) were recruited from Beijing Normal University (BNU) and were randomly allocated to the control ($n = 31$) or the intervention ($n = 32$) group. All participants were native Chinese speakers with little prior knowledge of the artists used in the current study. They reported normal or corrected-to-normal vision, provided informed consent, were individually tested in a sound-proofed cubicle, and received monetary reward for their participation. The study was approved by the Ethics Committee of the Faculty of Psychology, Beijing Normal University.

Materials

The experiment comprised two learning tasks, both of which consisted of a study, distractor, and test phase. In the pre-intervention task, 120 paintings by 12 relatively unfamiliar artists (Georges Braque, Henri-Edmond Cross, Judy Hawkins, Philip Juras, Ryan Lewis, Marilyn Mylrea, Bruno Pessani, Ron Schlorff, Georges Seurat, Ciprian Stratulat, George Wexler, and YieMei) were taken from Kornell and Bjork (2008), with 10 paintings by each artist. Six randomly selected paintings by each artist were used for learning and the remaining four for testing. The artists were distributed into two sets for interleaved and blocked study, with six artists in each set. Learning difficulty was matched between these two sets according to Yan et al. (2016). The assignment of artists to the interleaved or blocked set was counterbalanced across participants. To further avoid any potential influence of familiarity with the artists' painting styles, artists' names were replaced by 12 common Chinese boys' names as in Sun et al. (2022).

The materials in the post-intervention task were 120 paintings from another 12 artists (Carla Bosch, David Grossmann, Wassily Kandinsky, Peder Mork Monsted, Grandma Moses, Roger Mühl, Georgia O'Keeffe, Pierre-Auguste Renoir, Henri Rousseau, Egon Schiele, Maurice Utrillo, and Guim Tio Zarraluki) taken from Sun et al. (2022), with 10 paintings from each artist. As described by Sun et al. (2022), all paintings were landscapes or skyscapes selected to be similar to those used by Kornell and Bjork (2008). Once again, six paintings were randomly selected for learning with the remaining four for testing. These artists' names were replaced by another 12 popular Chinese boys' names.

All paintings were cropped to remove identifying characteristics (e.g., signatures) and resized to 900 × 750 pixels. The study materials are publicly available at OSF (https://osf.io/9w36k/). All stimuli were displayed via MATLAB *Psychtoolbox*.

Design and procedure

Experiment 1 involved a 2 (study schedule: blocked vs. interleaved) × 2 (group: control vs. intervention) mixed design with study schedule as a within-subjects factor and group as a between-subjects factor. The procedure in Experiment 1 was the same as in Sun et al.'s (2022) Experiment 1, except that the study pace of the post-intervention learning task was self-regulated: on each trial, participants chose which artist's painting they wanted to study next.

Figure 1A depicts the procedure. In the study phase of the pre-intervention task, participants viewed 72 paintings from 12 artists. Paintings from six artists were presented in a blocked manner while those from the other 6 artists were presented in an interleaved manner. For blocks with paintings shown in the blocked (B) format, six paintings from a given artist were shown one-by-one consecutively in random order (i.e., the interleaving distance is 0 out of 5). The order of six blocked artists was also randomly determined by the computer. For blocks with paintings shown in the interleaved (I) format, six paintings from six different artists were interspersed, ensuring that paintings by a given artist were separated by an average of five paintings from the other five artists (i.e., the mean interleaving distance is 5 out of 5). There were 12 blocks in total and the order of blocks was B I I B B I I B B I I B as in Kornell and Bjork (2008), with B representing blocked and I representing interleaved. Each painting, preceded by a fixation cross for 0.5s, was studied for 5s with the corresponding artist's name presented below it.

After the study phase, participants calculated three-digit additions and subtractions for 15s as a distractor task, and then took an induction test. Forty-eight new paintings from the same 12 artists (four paintings from each) were presented one-by-one in random order. In each test trial, following a fixation cross (0.5s), a painting was shown on the top of the screen with 12 buttons representing the artists' names shown below. Participants judged which artist was responsible for the painting by clicking the

corresponding button with the mouse. The induction test was self-paced without time pressure or feedback.

After completing the pre-intervention task, both groups reported which strategy they thought helped them learn better by selecting one of three options: *1. Blocking; 2. About the same; 3. Interleaving*. The definition of these two strategies had already been elaborated at the beginning of the experiment. After making this metacognitive judgement, they read the intervention instructions.

For the intervention group, participants were given details of their actual test performance in the pre-intervention task, as well as a summary of Kornell and Bjork's (2008) findings concerning the interleaving effect. The instructions were adapted according to each participant's performance. Two bar plots depicting the participant's own performance and the results of Kornell and Bjork's (2008) Experiment 1b were shown below the instructions (see Figure 1B). In contrast, participants in the control group read irrelevant instructions about the effects of emotion on memory, without reference to interleaving. A schematic illustration depicting the results from Kensinger and Corkin (2003) was presented below the instructions (see Figure 1C). The exact wording of intervention and control instructions is provided in the Supplementary Materials (SM).

To test the effect of the intervention on metacognitive awareness, after the intervention phase participants were again asked to report which strategy they thought was more efficient with the same three options: *1. Blocking; 2. About the same; 3. Interleaving*.

Then, participants in both groups engaged in the post-intervention task. As shown in Figure 2, the names of the 12 new artists were shown at the center of the screen, aligned in two rows. As soon as participants clicked the button representing the artist they wanted to study, a fixation cross appeared at the center of the screen for 0.5s, followed by a randomly selected painting from the chosen artist presented

for 5s. Then the program automatically returned to the selection interface. Once all 6 paintings of a given artist had been studied, the color of that artist's name turned from black to grey, indicating that there were no more paintings to study from that artist.

Following the study phase of the post-intervention task, participants engaged in a 15s calculation task. Then the test phase started, in which participants classified 48 new paintings from the 12 artists studied in the post-intervention task. Finally, participants were debriefed and thanked. The entire experiment lasted approximately 40 minutes.

Coding of study sequence metrics during the post-intervention task

Three indicators were calculated to quantify study sequence choices in the post-intervention task for both the control and intervention groups, including (1) interleaving frequency (i.e., the proportion of switching between different categories), (2) percentage of interleavers (i.e., proportion of participants who used the interleaving strategy more frequently than the blocked strategy), and (3) interleaving distance (i.e., mean number of inserted exemplars from other categories between two successive exemplars from the same category).

Interleaving frequency. In the self-regulated learning phase of the post-intervention task, there were 72 paintings to be studied. Thus, each participant had 71 opportunities to choose whether they would like to study a new exemplar from the same (blocked) or a different (interleaved) category after studying each prior exemplar. By the time a given participant finished learning the last exemplar of a category (i.e., the sixth painting of a certain artist), there were no remaining exemplars to be studied in that category and they had to shift to another category. Switches under such situations were compulsory and not actively chosen by participants themselves. Among the 12 categories, there were 11 obligatory switches in total. Therefore, only the remaining 60 choices (= 71-11) were truly discretionary.

Each of these 60 choices was either a switch between categories (interleaving) or remaining in the same category (blocking), and the proportion of interleaving (i.e., interleaving frequency) was calculated by dividing the number of times a participant chose to switch to a different category by the total number of choices (60). This calculation returns a value ranging from 0 (no interleaving) to 100% (full interleaving), representing a given participant's interleaving frequency.

To enable comparisons between studies, it is essential to explicitly detail the distinctions and connections between "interleaving frequency" used in the current study and similar metrics used in other studies. We classified each sequence choice as "switch" or "stay" depending on whether the category chosen in the current trial (j) was identical to the category chosen in the prior trial (j - 1). In contrast, Tauber et al. (2013) and Kornell and Vaughn (2018) calculated blocking rate following the criterion that Trial j was deemed blocked if the same category was chosen in either the prior (j - 1) or the subsequent trial (j + 1); otherwise it was categorized as interleaved. To clarify the differences between our coding method and theirs, take the sequence AABB as an example. We would count Trials 2 and 4 as "blocked" and Trial 3 as "interleaved". By contrast, Tauber et al. (2013) and Kornell and Vaughn (2018) would count all four trials as "blocked".

In fact, our calculation of interleaving frequency closely mirrored Lu et al.'s (2021) approach to compute the proportion of switching between different categories. The key difference lies in our exclusion of all obligatory switches (11) from the total number of choices (71), so that the calculated value of interleaving frequency ranges from 0 (pure blocking) to 100% (pure interleaving). Lu and colleagues, on the other hand, included obligatory switches in their analysis.

Percentage of interleavers. Following Tauber et al. (2013), participants whose interleaving frequency exceeded 50% were classified as interleavers whereas those whose proportion of interleaving

frequency was equal to or lower than 50% were classified as blockers.

Interleaving distance. Switching frequency reflects the frequency with which participants chose to switch to a different category, but it does not represent the distance between exemplars from the same category. As discussed above, large interleaving is more effective for enhancing inductive learning than small interleaving. Below, we quantified the degree of interleaving by averaging the mean distance between exemplars from the same category.

In alignment with the operationalization of exemplar-to-exemplar spacing adopted by Kornell and Vaughn (2018), interleaving distance in the current study is defined as the number of exemplars sandwiched between two adjacent exemplars from the same category. For example, if a painting by Carla was selected in the first trial followed by Carla's second painting in the fifth trial, this artist's two exemplars were separated by three artworks from other artists, thus establishing an interleaving distance of 3. All six paintings belonging to the same artist were selected from the study sequence, and hence there were five distance values for each artist (category). The mean of these five distances was calculated as the mean interleaving distance for this category. Then, the mean distances of all 12 categories were averaged for each participant. Therefore, an average distance of 0 indicates no interleaving distance, with 11 indicating maximal interleaving distance. Furthermore, it should be noted that frequent switching between categories does not necessarily yield a large interleaving distance as switches could occur in a confined number of categories (see the example described in the Introduction).

Transparency and Openness

The determination of our sample size, intervention methods, and outcome measures are described in the above section. Exact wording of the intervention and placebo instructions is reported in the SM. All data and materials of artists' paintings (used in Experiments1-3) and butterfly species (used in

Experiments 2 & 3) are publicly available at https://osf.io/9w36k/. All data were analyzed using JASP 0.16.2 (JASP Team, 2022). The experiment was not preregistered.

Results

Bayes factors (BF_{01} or BF_{excl}) are reported when the conventional frequentist *t* or *F* tests do not reach statistical significance at the $\alpha = .05$ level. BF_{01} was used to quantify the relative evidence in favor of the null hypothesis over the alternative (Jeffreys, 1961; Kass & Raftery, 1995). A chi-squared test for independence was used to investigate whether metacognitive judgements differed significantly between the control and intervention groups. A chi-squared test for goodness-of-fit was applied to determine whether participants demonstrated a preference for any one of the three options (i.e., *blocking*, *about the same*, and *interleaving*).

The main text elucidates the key findings concerning the intervention effects on metacognitive awareness, interleaving strategy usage, and classification accuracy. Other exploratory analyses assessing (1) strategy shifts throughout self-regulated learning, (2) how self-usage of the interleaving strategy affects classification performance, (3) the influence of personalized feedback in Experiments 1 and 2, (4) a mediation analysis examining the mediating role of self-regulated strategy usage in the intervention effect on classification performance in Experiment 3 are reported in the SM.

Test performance in the pre-intervention task

Figure 3A shows test performance in the pre-intervention task in each of the control and intervention groups. A 2 (study schedule: blocked vs. interleaved) × 2 (group: control vs. intervention) mixed analysis of variance (ANOVA) was conducted with study schedule as a within-subjects factor and group as a between-subjects factor. The results showed a main effect of study schedule, $F(1, 61) = 45.44$, $p < .001$, $\eta^2_p = .43$, with superior classification accuracy for interleaved than for blocked artists and reflecting an

interleaving effect. Neither the effect of group, $F(1, 61) = 2.05$, $p = .16$, $\eta_p^2 = .03$, $BF_{excl} = 1.45$, nor the interaction between schedule and group was statistically significant, $F(1, 61) = 0.09$, $p = .77$, $\eta_p^2 = .001$, $BF_{excl} = 3.79$. Pre-planned paired *t*-tests showed that the interleaving effect was significant in both the control, $t(30) = 5.49$, $p < .001$, $d = 0.99$, and intervention groups, $t(31) = 4.38$, $p < .001$, $d = 0.77$. The classic interleaving effect originally documented by Kornell and Bjork (2008) was evidently replicated here.

Metacognitive awareness before intervention

Descriptive data relating to metacognitive judgements made before the intervention phase are reported in Table 1. As expected, there was no significant difference in judgments between the two groups before the intervention phase, $p = .94$.[1] Therefore, we collapsed the data across the two groups to increase statistical power for further analyses.

A chi-squared test revealed that participants' preference towards the two learning strategies differed significantly, $\chi^2(2) = 20.10$, $p < .001$. Confirming the well-established metacognitive illusion, the percentage (57%) of participants believing that blocking is superior to interleaving (B > I) was greater than the percentage (32%) believing B < I, $\chi^2(1) = 4.57$, $p = .03$, and 11% of participants believed the two strategies to be about equally effective (B = I).

Metacognitive awareness after intervention

After the intervention phase, a significant difference in participants' efficacy judgments was detected between the two groups, $p = .04$ (see Figure 3B). Thus, the results of the two groups are reported separately.

[1] The Fisher-Freeman-Halton's exact test was employed when the assumptions for chi-square test were not met, specifically when the expected count(s) in one or more cells is smaller than five. This situation arose primarily due to the low number of participants selecting the "about the same" option. Note that the test does not generate a test statistic, so in these instances only *p*-values are reported.

For participants in the control group, their judgment choices across the three options differed significantly, $\chi^2(2) = 8.19$, $p = .02$. The proportion of participants believing B > I (55%) was numerically greater than that believing B < I (32%), $\chi^2(1) = 1.82$, $p = .18$. The remaining 15% of participants believed B = I. In the intervention group, participants' choices across the three options also differed significantly, $\chi^2(2) = 17.31$, $p < .001$. A majority of participants now believed B < I (66%), greater than the percentage believing B > I (28%), $\chi^2(1) = 4.80$, $p = .03$, and there were 6% of participants who believed B = I.

To summarize, reading control instructions failed to change participants' metacognitive illusion about the benefits of interleaving, whereas providing feedback on each participant's own performance and the findings of Kornell and Bjork (2008) enhanced individuals' appreciation of the advantages of interleaving.

Interleaving usage in the post-intervention task

Three measures (i.e., interleaving frequency, proportion of interleavers, and interleaving frequency) of self-usage of the interleaving strategy were calculated. Detailed results of these measures are provided in Table 2. An independent-samples *t* test showed that interleaving frequency was higher in the intervention group ($M = 56\%$, $SD = 31\%$) than that in the control group ($M = 39\%$, $SD = 33\%$), difference = 17%, [0.5%, 33%], $t(61) = 2.06$, $p = .04$, $d = 0.52$, reflecting that correcting metacognitive bias via the instruction intervention promotes interleaving frequency during self-regulated learning. As presented in Table 2, the proportion of interleavers in the intervention group (56%) was significantly higher than that in the control group (29%), $\chi^2(1) = 4.76$, $p = .03$, confirming that instructions about the benefits of interleaving enhance self-usage of the interleaving strategy. However, although in the predicted direction, there was no statistically detectable difference in interleaving distance between the intervention ($M = 6.26$ out of 11, $SD = 3.64$) and control ($M = 5.41$ out of 11, $SD = 3.49$) groups, difference = 0.84, [-0.96, 2.64],

$t(61) = 0.94$, $p = .35$, $d = 0.24$, $BF_{01} = 2.69$.

Overall, although instructions about the merits of interleaving promoted interleaving frequency and enhanced the proportion of participants who used the interleaving strategy more frequently than the blocked strategy, they did not stimulate participants to separate exemplars from the same category by larger distances.

Test performance in the post-intervention task

Table 2 shows final classification performance in the post-intervention task. There was no statistically detectable difference in classification accuracy between the two groups, difference = 4%, [-8%, 15%], $t(61) = 0.61$, $p = .55$, $d = 0.18$, $BF_{01} = 3.33$, although classification accuracy was numerically better in the intervention ($M = 72\%$, $SD = 24\%$) than in the control group ($M = 68\%$, $SD = 22\%$).

Discussion

In Experiment 1, the typical interleaving effect identified by Kornell and Bjork (2008) was successfully replicated using the same painting materials. We also confirmed Sun et al.'s (2022) findings that an instruction intervention – comprised of overall feedback on each participant's own classification performance and a summary of the general results regarding the interleaving effect from previous studies – was effective in alleviating participants' metacognitive illusions. In the self-regulated learning phase, the instructional intervention altered strategy employment by promoting switches between categories rather than sticking to a specific category. Furthermore, the proportion of interleavers was increased by the instructional intervention. Since frequent between-category switching occurred among a restricted number of categories, the interleaving distance did not reliably change. This outcome is not surprising, as the instructions primarily focused on the benefits of interleaving, without specific emphasis on the benefits of large over small interleaving. Such guidance encouraged participants to alternate among

different categories, rather than attempting to alternate among a large number of categories. Consequently, the instruction intervention did not improve post-intervention classification accuracy.

Experiment 2

Experiment 2 had three objectives. Firstly, it aimed to replicate the findings of Experiment 1. Secondly, it investigated whether the effect of the instruction intervention transfers from learning artists' painting styles to a completely different domain, that is, learning butterfly species. Note that Sun et al. (2022) recently reported that this intervention effect is transferable. Thirdly, though the instructional intervention was not tailored to elicit extensive interleaving, the small number of categories (i.e., only 12 artists) in Experiment 1's post-intervention task might be insufficient to manifest an intervention effect on interleaving distance. This in turn resulted in minimal difference in classification accuracy between the two groups. Therefore, in Experiment 2, 16 species of butterflies were used in the post-intervention task, allowing participants to interleave between more categories, which might yield detectable differences in interleaving distance and induction performance between the groups.

Method

Participants

Based on the same plan as Experiment 1, the sample size was set to 30 participants per group. Finally, 68 participants (*M* age = 20.99, *SD* = 2.17; 96% female) were recruited from BNU and randomly allocated to the control (n = 34) or intervention (n = 34) groups. All participants were Chinese with little prior knowledge about the artists or butterfly species used in the current experiment. They reported normal or corrected-to-normal vision, provided informed consent, were tested individually in a sound-proofed cubicle, and received monetary reward for their participation.

Materials

The materials used in the pre-intervention task were identical to those used in the equivalent phase of Experiment 1. Stimuli in the post-intervention task were 80 images of 16 butterfly species (Admiral, American, Baltimore, Cooper, Eastern Tiger, Hairstreak, Harvester, Mark, Painted Lady, Pine Elfin, Pipevine, Sprite, Tipper, Tree Satyr, Viceroy, and Wood Nymph) taken from Birnbaum et al. (2013), with five images from each species. Four randomly selected images of each species were used for learning and the fifth for testing. The butterfly names were translated into Chinese and the images were readjusted to prevent any physical characteristics implied by the names. All images were resized to 900 × 750 pixels. The resized images are publicly available at OSF (https://osf.io/9w36k/).

Design and procedure

The design and procedure of Experiment 2 were the same as those of Experiment 1 with the exception that learning painting styles was replaced by learning butterfly species in the post-intervention task. In the self-regulated learning phase of the post-intervention task, 16 buttons corresponding to the species were shown at the center of the screen in two rows. Once participants had clicked the button representing the species they wanted to study next, a fixation cross appeared at the center of the screen for 0.5s, followed by a randomly selected butterfly image from the chosen species presented for 5s.

After studying 64 butterfly images, participants engaged in a 15s calculation task as distraction. In the test phase, 16 new butterfly images (one from each species) were shown one-by-one at the top of the screen in a random order. Below the image were 16 buttons, each tagged with the name of a particular butterfly species, aligned in two rows. Participants clicked the button representing the species of the current butterfly image without time pressure or feedback. They were debriefed and thanked at the end of the post-intervention task. The entire experiment lasted approximately 40 minutes.

Results

Test performance in the pre-intervention task

Figure 4A shows test performance in the pre-intervention task in the control and intervention groups. A mixed ANOVA manifested a main effect of study schedule, $F(1, 66) = 96.18$, $p < .001$, $\eta_p^2 = .59$. There was no effect of group, $F(1, 66) = 2.47$, $p = .12$, $\eta_p^2 = .04$, $BF_{excl} = 1.33$, nor any interaction between study schedule and group, $F(1, 66) = 0.22$, $p = .64$, $\eta_p^2 = .003$, $BF_{excl} = 3.53$. Pre-planned paired *t*-tests showed that the interleaving effect was significant in both the control, $t(33) = 6.84$, $p < .001$, $d = 1.17$, and intervention groups, $t(33) = 7.08$, $p < .001$, $d = 1.22$.

Metacognitive awareness before intervention

Descriptive data relating to metacognitive judgements measured before the intervention phase are presented in Table 1. Metacognitive awareness did not differ between the two groups, $p = .46$, so we collapsed the data across the two groups to increase power for further analyses.

There was a significant difference in participants' preferences for the two strategies, $\chi^2(2) = 25.53$, $p < .001$. The percentage of participants believing B > I (59%) was greater than the percentage believing B < I (32%), $\chi^2(1) = 5.23$, $p = .02$, and 9% of participants believed B = I.

Metacognitive awareness after intervention

Figure 4B presents metacognitive judgments made after the intervention phase. There was a significant difference in beliefs between the two groups, $p = .001$. Hence, the results in the two groups are reported separately.

For participants in the control group, metacognitive bias regarding the effectiveness of blocking was maintained, $\chi^2(2) = 9.94$, $p = .007$. Numerically more participants believed B > I (56%) than B < I (32%), $\chi^2(1) = 2.13$, $p = .14$. The percentage believing B = I was 12%. For participants in the intervention group,

their choices across the three options also differed significantly, $\chi^2(2) = 29.18$, $p < .001$. Specifically, a majority now believed B < I (77%), greater than the percentage believing B > I (18%), $\chi^2(1) = 12.50$, $p = .03$, and 6% of participants believed B = I.

Interleaving usage in the post-intervention task

Table 2 reports the proxy measures of self-employment of interleaving. Since participants studied four butterfly images from each of the 16 species, the total number of study trials was 64. Each participant decided which species to study on 63 trials in total. After removing 15 compulsory switches, the number of authentic discretionary choices was 48, which was used as the denominator to calculate interleaving frequency.

As shown in Table 2, participants in the intervention group ($M = 64\%$, $SD = 33\%$) interleaved category exemplars more frequently than those in the control group ($M = 42\%$, $SD = 29\%$), difference = 22%, [7%, 37%], $t(66) = 2.97$, $p = .004$, $d = 0.72$, replicating the main finding of Experiment 1. Furthermore, the percentage of interleavers in the intervention group (71%) was substantially greater than in the control group (32%), $\chi^2(1) = 9.95$, $p = .002$, re-confirming the instruction intervention effect on interleaving preference. Consistent with Experiment 1, an independent-samples *t* test again showed no detectable difference in interleaving distance between the intervention ($M = 8.81$ out of 15, $SD = 4.79$) and control ($M = 7.48$ out of 15, $SD = 4.44$) groups, difference = 1.33, [-0.91, 3.57], $t(66) = 1.19$, $p = .24$, $d = 0.29$, $BF_{01} = 2.21$.

Test performance in the post-intervention task

The proportion of correctly classified butterflies did not differ significantly between the intervention ($M = 39\%$, $SD = 14\%$) and control ($M = 41.0\%$, $SD = 13.9\%$) groups, difference = -2%, [-9%, 5%], $t(66) = -0.47$, $p = .63$, $d = -0.14$, $BF_{01} = 3.63$ (see Table 2). The numerically reversed pattern of results in

relation to the *a priori* prediction might derive from sampling or measurement error as there were only 16 test trials in the post-intervention task. In brief, aligned with Experiment 1, Experiment 2 again observed minimal influence of the instruction intervention on inductive learning.

Discussion

Experiment 2 conceptually replicated the main findings of Experiment 1. In addition, it further demonstrated that the intervention effect on interleaving frequency transferred to learning of categories in a different domain (for related findings, see Sun et al., 2022). However, the intervention had minimal impact on interleaving distance or classification accuracy, even when there were a larger number of categories studied in the post-intervention task.

Experiment 3

The first two experiments jointly documented that instructions consisting of individualized feedback on blocked vs. interleaved performance and a summary of previous findings about the interleaving effect promoted self-usage of the interleaving strategy (i.e., interleaving frequency and percentage of interleavers). Experiment 2 further confirmed that the intervention effect on interleaving frequency transferred from learning artists' painting styles to learning butterfly species. However, the intervention failed to increase interleaving distance or classification accuracy, even when there were a larger number of categories in the self-regulated learning task. Leveraging the strengths of interleaving and spacing, larger interleaving distances should be more beneficial for inductive learning than small distances. Indeed, Experiments 1 and 2 consistently observed a positive correlation between interleaving distance and classification accuracy (see the SM for detailed results). The null intervention effect on interleaving distance may explain why the instruction intervention produced minimal potentiating effect on classification performance.

Going beyond Experiments 1 and 2, Experiment 3 developed new intervention instructions to increase both interleaving frequency and distance, exploring if such instructions would motivate participants to interleave exemplars with larger distances and hence enhance inductive learning. To achieve these aims, three sets of artists were studied in (1) blocked, (2) small interleaved, and (3) large interleaved conditions, respectively, in the pre-intervention task. The instruction interventions additionally incorporated the findings of Birnbaum et al.'s (2013) Experiment 3 as evidence demonstrating the efficacy of large compared to small interleaving, along with individual performance feedback from the pre-intervention task.

In Experiment 2, the number of study (64 trials with four butterflies from each species) and test (16 trials with one new butterfly image from each species) trials in the post-intervention task might be too small to produce a detectable difference between the two groups. Therefore, in Experiment 3, the number of exemplars for each species was increased from 5 to 10, with 6 used for studying (96 trials in total) and the remaining 4 for testing (64 trials in total).

Method

Participants

The sample size was pre-determined in the same way as Experiments 1 and 2. Ultimately, 66 participants (*M* age = 21.14, *SD* = 1.91; 91% female) were recruited from BNU and randomly assigned to the control (n = 33) or intervention (n = 33) groups. All were native Chinese speakers, had normal or corrected-to-normal vision, lacked familiarity with the artists and butterfly species used in the present experiment, signed informed consent, were individually tested in a sound-proofed cubicle, and received monetary compensation after the experiment.

Materials

Experiment 3 contained two learning tasks. Paintings by 12 artists from Kornell and Bjork (2008) were supplemented with paintings by six additional artists (Carla Bosch, Roger Mühl, Georgia O' Keeffe, Pierre-Auguste Renoir, Henri Rousseau, and Guim Tio Zarraluki) from Sun et al. (2022) as stimuli in the pre-intervention task. Each artist had 10 paintings, with six used for studying and the remaining four for testing. The 18 artists were evenly assigned to three sets, with minimal difference in learning difficulty among the sets according to the classification accuracy of paintings by each artist in a pilot study. Paintings in these three sets were studied under different schedules (blocked vs. small interleaved vs. large interleaved). Assignment of sets to conditions was counterbalanced across participants. Again, all artists' names were replaced by 18 popular Chinese boys' names.

In the post-intervention task, 16 butterfly species were taken from Birnbaum et al. (2013) with five images from each species. To increase the number of study and test trials, five new butterfly images from each species were included. Therefore, in Experiment 3 each species was represented by 10 exemplars, with six used for studying and the remaining four for testing. All paintings and butterfly images were resized to 900 × 750 pixels.

Design and Procedure

Experiment 3 employed a 3 (study schedule: blocked vs. small interleaved vs. large interleaved) × 2 (group: control vs. intervention) mixed factorial design, with study schedule manipulated within-subjects and group manipulated between-subjects.

In the study phase of the pre-intervention task, participants were presented 108 paintings consecutively, with six by each artist. Paintings by six artists were shown in a blocked (B) schedule, with paintings by another six artists shown in a small interleaved (SI) schedule and those by the remaining six

artists in a large interleaved (LI) schedule. The study schedules were organized as follows: B B SI SI LI LI B B SI SI LI LI B B SI SI LI LI. Within each B B sequence, six paintings by the same artist were presented consecutively followed by six paintings by another artist, resulting in an interleaving distance of 0 out of 5. Within each SI SI sequence, paintings by two artists alternated until all 12 paintings had been studied, resulting in an interleaving distance of 1 out of 5. Within each LI LI sequence, six paintings by six different artists were interspersed and presented in succession, resulting in an interleaving distance of 5 out of 5. The order of artists studied in each of the schedules was randomized anew for each participant, as well as the painting order of each artist. To avoid inconsistent retention intervals for different study schedules (i.e., blocked artists are always studied first followed by SI and LI artists), the order of these schedules was counterbalanced across participants.

After the study phase, participants calculated three-digit additions and subtractions for 15s and then engaged in the induction test. Seventy-two new paintings by the 18 studied artists (four paintings by each) were presented one-by-one in a random order. In each test trial, after a 0.5s fixation, a painting was displayed at the top of the screen with 18 buttons labelling artists shown below, and participants selected the artist responsible for the painting by clicking the corresponding button with the mouse.

After completion of the pre-intervention task, participants were asked to indicate which strategy they thought was the most effective for facilitating their learning by selecting one from three options: *1. Blocking; 2. Small interleaving; 3. Large interleaving*. Then they proceeded to the intervention phase. For participants in the intervention group, they were first provided their test performance in the pre-intervention task, followed by the results from Kornell and Bjork's (2008) Experiment 1a to demonstrate the interleaving effect. More importantly, Birnbaum et al.'s (2013) Experiment 3 was briefly introduced

with emphasis on the finding that large interleaving is more effective than small interleaving.[2] The exact wording of intervention instructions can be found in the SM.

For participants in the control group, they read the same control instructions as used in Experiments 1 and 2. To examine the intervention effect on metacognitive awareness, participants were again asked to indicate which strategy they thought was most effective by selecting one from three options: *1. Blocking; 2. Small interleaving; 3. Large interleaving*. Next, both groups completed the post-intervention task. This was identical to that in Experiment 2 except the number of study trials increased from 64 to 96 (six butterfly images from each species) and the number of test trials increased from 16 to 64 (four butterfly images from each species).

Results

Test performance in the pre-intervention task

Test performance in the pre-intervention task was analyzed with a 3 (study schedule: blocked vs. small interleaved vs. large interleaved) × 2 (group: control vs. intervention) mixed ANOVA with study schedule as the within-subjects variable and group as the between-subjects variable. As shown in Figure 5A, there was a main effect of study schedule, $F(2, 128) = 49.48$, $p < .001$, $\eta_p^2 = .44$, but no main effect of group, $F(1, 64) = 0.78$, $p = .38$, $\eta_p^2 = .01$, $BF_{excl} = 2.77$, nor interaction between study schedule and group, $F(2, 128) = 0.25$, $p = .78$, $\eta_p^2 = .004$, $BF_{excl} = 8.45$.

Pairwise *t* tests suggested that participants correctly classified roughly equal numbers of paintings in the blocked ($M = 31\%$, $SD = 19\%$) and small interleaved ($M = 33\%$, $SD = 20\%$) conditions, $t(65) = -0.94$, $p = .35$, $d = -0.12$, $BF_{01} = 4.84$. Since small interleaving in Experiment 3 only encompassed switching

[2] Note that Birnbaum et al. (2013) intended to examine the effects of temporal spacing inherent in interleaving, so they termed separating exemplars from a given category by 3 exemplars from other categories as small spacing, and by 15 exemplars from other categories as large spacing. Essentially, their small spacing is small interleaving among 4 categories whereas large spacing is large interleaving among 16 categories. Therefore, we used the term "interleaving" rather than "spacing" in the instructions to prevent ambiguity.

between two artists, it is conceivable that the discriminative-contrast benefits of interleaving did not come into full play. Classification accuracy in the large interleaved condition (M = 54%, SD = 20%) was significantly higher than in both the blocked, $t(65) = 7.92$, $p < .001$, $d = 0.97$, and small interleaved conditions, $t(65) = 9.08$, $p < .001$, $d = 1.12$. Overall, these findings successfully replicate the advantage of large versus small interleaving and blocking (Birnbaum et al., 2013).

Metacognitive awareness before intervention

Metacognitive judgements measured before and after the intervention phase are depicted in Figure 5B. Before the intervention phase, a bulk of participants believed that blocking was the most effective strategy for inductive learning. A chi-squared test showed that metacognitive awareness did not differ between the groups, $\chi^2(2) = 2.23$, $p = .33$. Therefore, judgment data were merged to increase statistical power for further analyses.

A chi-squared goodness of fit test revealed that the number of participants favoring each of the three study strategies was not evenly distributed, $\chi^2(2) = 19.91$, $p < .001$. A majority believed blocking was optimal (59%), greater than the percentage favoring small interleaving (23%), $\chi^2(1) = 10.67$, $p = .001$, and the percentage favoring large interleaving (18%), $\chi^2(1) = 14.29$, $p < .001$. There was no significant difference between the percentages believing small and large interleaving was the most efficient strategy, $\chi^2(1) = 0.33$, $p = .56$. Overall, over 80% of participants lacked awareness that large interleaving is the best strategy.

Metacognitive awareness after intervention

Metacognitive judgments made after the intervention phase differed significantly between the two groups, $\chi^2(2) = 15.02$, $p = .001$. Thus, the results of the two groups are reported separately (see Table 1). In the control group, although the percentage of participants choosing each of three strategies did not

differ significantly, $\chi^2(2) = 2.91$, $p = .23$, the percentage favoring blocking (46%) was numerically higher than the percentages favoring small (33%) and large interleaving (21%). By contrast, in the intervention group, choices across the three options differed significantly, $\chi^2(2) = 16.91$, $p < .001$. Two-thirds of participants (67%) now correctly judged large interleaving to be optimal, greater than both the percentage (12%) believing blocking to be optimal, $\chi^2(1) = 12.46$, $p < .001$, and than the percentage (21%) believing small interleaving to be optimal, $\chi^2(1) = 7.76$, $p = .005$. There was little difference between the percentages favoring blocking and small interleaving, $\chi^2(1) = 0.82$, $p = .37$.

To sum up, these results confirm that control instructions had little impact on participants' appreciation of the effectiveness of large interleaving, whereas intervention instructions highlighting the benefits of interleaving and large interleaving successfully reversed their metacognitive bias.

Interleaving usage in the post-intervention task

Table 2 details the indicative measures that evaluate self-utilization of the interleaving strategy in the post-intervention task. With 96 butterfly images (6 images of each of 16 species) learned in the study phase of the post-intervention task, each participant had 95 opportunities to choose between switching and staying. After removing 15 obligatory switches between 16 categories, the remaining 80 choices (= 95-15) were discretionary. The percentage of switch choices quantified interleaving frequency. An independent-samples *t* test showed that participants in the intervention group ($M = 68\%$, $SD = 33\%$) switched more frequently between categories than those in the control group ($M = 40\%$, $SD = 25\%$), difference = 27%, [13%, 42%], $t(64) = 3.80$, $p < .001$, $d = 0.94$. In addition, the instruction intervention resulted in a significant increase in the percentage of interleavers from 39% to 70% across the control and intervention groups, $\chi^2(1) = 6.11$, $p = .01$.

Of critical interest, interleaving distance in the intervention group ($M = 9.89$ out of 15, $SD = 3.61$)

was significantly larger than that in the control group (M = 7.22 out of 15, SD = 4.78), difference = 2.67, [0.59, 4.75], $t(59.53) = 2.57$, $p = .01$, $d = 0.63$, reflecting that adding the benefits of large interleaving to the instruction intervention motivated participants to interleave category exemplars at larger distances (see Table 2).

Test performance in the post-intervention task

As shown in Table 2, participants in the intervention group correctly classified more butterfly images (M = 57%, SD = 11%) than those in the control group (M = 50%, SD = 16%), difference = 7%, [0.3%, 14%], $t(64) = 2.07$, $p = .04$, $d = 0.51$, reflecting an intervention effect on inductive learning.

Discussion

Experiment 3 provided evidence that switching between only a small number of categories yields minimal enhancement on inductive learning, while switching among a large number of categories produces a reliable benefit. Furthermore, the intervention instructions highlighting the benefits of interleaving (over blocking) and the merits of large interleaving (over small interleaving) were successful in overcoming participants' erroneous beliefs about the most effective strategy, promoting interleaving frequency, increasing both the percentage of interleavers and the magnitude of their chosen interleaving distances, and most importantly, producing better inductive learning outcomes. Again, the intervention effect on strategy employment transferred from learning painting styles to learning butterfly species.

Exploratory Analyses

In addition to the results reported above, exploratory analyses were conducted for each of Experiments 1-3 to examine (1) how participants' interleaving frequency varied across the study phase in the post-intervention task, (2) the relationship between interleaving frequency and classification performance in the post-intervention task, and (3) the relationship between interleaving distance and

classification performance in the post-intervention task. The detailed results are reported in the SM.

In brief, Experiments 1-3 consistently observed that in both the control and intervention groups, interleaving frequency changed over the course of self-regulated learning: participants tended to interleave at the beginning of the study phase but steadily converted to blocking near the end (for related findings, see Lu et al., 2021). This is consistent with the idea that at the beginning of the study phase, learners scanned across categories for an overview of the to-be-learned materials, but then focused on each category separately (Lu et al., 2021). Furthermore, both interleaving frequency and interleaving distance served as potential predictors of classification performance as they positively correlated with classification accuracy in the post-intervention task in each of Experiments 1-3. These findings provide insights into how to improve the effectiveness of interventions designed to facilitate inductive learning: switching more frequently between different categories and more importantly, enlarging interleaving distance between exemplars from the same category.

As Experiment 3 revealed a significant intervention effect on classification performance in the post-intervention task, a mediation analysis was conducted to investigate whether this effect was mediated by the use of the interleaving strategy during self-regulated learning phase. The result of a complete mediation suggested that instructional intervention enhanced self-employment of the interleaving strategy, which in turn led to improved learning outcomes after the intervention.

General Discussion

In the wake of the discovery of the interleaving effect by Kornell and Bjork (2008), this effect has been broadly verified across different types of materials and study-test intervals (see the Introduction). Experiments 1 and 2 successfully replicated the classic interleaving effect using the same materials and procedure as Kornell and Bjork (2008). Typically, learners reap the benefits of interleaving versus

blocking. However, the magnitude of the interleaving effect is contingent on interleaving distance. Birnbaum et al. (2013) compared classification accuracy between small and large interleaving conditions and demonstrated that large interleaving (i.e., interleaving among a large number of categories) is more beneficial for inductive learning than small interleaving (i.e., interleaving among a small number of categories). Consistent with this, Experiment 3 observed that interleaving between only two categories yielded minimal enhancement compared to pure blocking (d = 0.12), whereas interleaving among six categories was far more beneficial than both small interleaving (d = 1.12) and pure blocking (d = 0.97), suggesting that interleaving distance is an important moderator of the interleaving effect.

Along with the robustness of the interleaving effect, research has also repeatedly documented learners' deep-rooted but mistaken belief that massing exemplars from the same category is superior to mixing exemplars from different categories (e.g., Guzman-Munoz, 2017; Kornell & Bjork, 2008; Kornell et al., 2010; Yan et al., 2016; Yan et al., 2017). Metacognitive unawareness of the benefits of interleaving then leads to underemployment of the interleaving strategy and thus produces suboptimal performance (e.g., Lu et al., 2021; Sun et al., 2022; Tauber et al., 2013). Indeed, in both Experiments 1 (68%) and 2 (68%), most participants did not appreciate the benefits of interleaving. Accordingly, they underemployed the interleaving strategy during the subsequent self-regulated learning tasks (as reflected by the high proportions of participants classified as "blockers"). Therefore, it is of considerable importance to explore methods to alleviate or reverse learners' metacognitive illusions, so that self-employment of interleaving can be promoted in the context of inductive learning.

Correcting metacognitive illusions and enhancing interleaving strategy use through explicit instructions

In a previous effort to correct such ubiquitous metacognitive illusions, personal experience of the

superiority of interleaving as well as elaborated information explaining the underpinnings of the interleaving effect failed to drive participants to appreciate the benefits of interleaving (Yan et al., 2016). Specifically, after participants undergoing both blocked and interleaved schedules, Yan et al. (2016) asked them which schedule they believed produced better learning of the artists' styles. Similar to the metacognitive judgements in the current study, participants were asked to choose among three choices: "blocking is better," "interleaving is better," and "about the same." Surprisingly, though Yan et al. explicitly informed participants of a counterintuitive fact that the vast majority (90%) benefit from interleaving and explained why interleaving is more effective for inductive learning, only 36%-52% of participants responded "interleaving is better" across their Experiments 3-5.

In contrast, incorporating individualized feedback on blocked vs. interleaved performance and a summary of the results from Kornell and Bjork (2008) successfully overcame the "blocking is superior to interleaving" illusion, as demonstrated by Sun et al. (2022), a finding also reproduced here. In our Experiments 1 and 2, 66-77% of participants in the intervention group came to acknowledge the merits of interleaving after being exposed to the instruction intervention, whereas, after the control intervention, only 31%-38% of participants perceived the benefits of interleaving. Furthermore, our Experiment 3 successfully convinced 67% of participants in the intervention group that large interleaving is more effective than small interleaving and pure blocking when being explicitly informed about the benefits of large interleaving in the intervention instructions. In contrast, after the control intervention, 79% of participants still mistakenly reported that the most effective strategy was blocking or small interleaving.

Several factors are available to account for the discrepancies between the present study and Yan et al.'s (2016) study regarding the effectiveness of interventions designated to enhance metacognitive awareness of the benefits of interleaving. In Yan et al.'s (2016) Experiment 5, paintings by blocked and

interleaved artists were shown inside distinctive frames during the study phase to help participants readily link the paintings to the schedules in which they were presented.[3] However, the presence of such salient frame-schedule mappings may have distracted participants' attention away from monitoring their strategy experience during learning, thus impeding their ability to accurately assess the effectiveness of different strategies. According to the monitoring assumption within the knowledge updating framework, learners must accurately monitor the differential effectiveness during learning or while being tested to acquire strategy knowledge from task experience (Dunlosky & Hertzog, 2000). Research has indicated that, when repeatedly asked about the perceived effort and learning associated with interleaved and blocked practice, participants were able to recognize, through on-task experience, that interleaved practice became less effortful over time and was more beneficial to their learning than blocked practice (Onan et al., 2022).

Another major difference between the current study and Yan et al.'s is how feedback was afforded. In the current experiments, participants received feedback on their overall classification performance in each study condition and could use this information as a basis for their second judgement of strategy effectiveness. Yan et al.'s (2016) Experiment 5, in contrast, provided participants with feedback on the accuracy of their responses along with the correct artist's name following each of their classification decisions in the induction test. It is evident that overall feedback for each study schedule conveys the merits of interleaving more straightforwardly than does trial-by-trial feedback.

Moreover, the phrasing of Yan et al.'s (2016) intervention instructions tended to imply that though many learners capitalize on intermixing exemplars, this is not the case for a minimal number of "uncommon" learners. McDaniel and Einstein (2020) suggested that learners would not put a strategy into

[3] Yan et al.'s (2016) study consisted of six experiments. We selected their Experiment 5 for comparison with our investigation into the intervention effects on metacognitive awareness of the interleaving effect, due to its resemblance to our pre-intervention task and the subsequent intervention phase. Additionally, their Experiment 5 implemented the most comprehensive intervention approaches across their six experiments, but these interventions still failed to induce a shift in participants' preference from blocking to interleaving.

effect unless they are convinced it works for *themselves*, and Yan et al.'s (2016) participants might identify themselves as one of the "special populations" for whom interleaving does not work. This particular bias towards individual uniqueness may be relatively mitigated in the current study as our instructions presented participants with a figure illustrating their own performance in both blocked and interleaved conditions, alongside the results showing the interleaving effect documented by Kornell and Bjork (2008). Hence, most participants could conveniently appreciate that their personalized pattern of performance was consistent to the pattern of interleaving effect in the averaged data of their peers, thereby becoming more readily persuaded that interleaved practice is more beneficial for them as well.

In light of the intimate relationship between strategy effectiveness beliefs and strategy usage during self-regulated learning (Bjork et al., 2013; Yang et al., 2017), we argue that enhanced metacognitive awareness of the superiority of interleaving is likely to correspondingly increase its use in the post-intervention task. As anticipated, in Experiments 1 and 2, participants in the intervention group chose to study 56-64% of exemplars via the interleaving strategy while the corresponding proportion was only 39-42% in the control group. Furthermore, 56-71% of participants in the intervention group switched on more than half the trials, in stark contrast to the 29-32% of such interleavers in the control group.

In addition to interleaving frequency and percentage of interleavers, the self-regulated learning task employed in the current study allows us to calculate a new proxy reflecting the degree to which exemplars are interspersed, namely interleaving distance. Based on both the discriminative-contrast (e.g., Kang & Pashler, 2012) and study-phase retrieval (e.g., Dunlosky et al., 2013) hypotheses, interleaving distance is theoretically a pivotal factor to optimize classification accuracy. Although interleaving frequency significantly correlated with classification accuracy in the post-intervention task (*r*s = .23-.44 across Experiments 1-3; see the SM), Experiments 1 and 2 found that merely enhancing interleaving frequency

was insufficient to yield an impact on classification accuracy. Due to comparable interleaving distance between the control and intervention groups in Experiments 1 and 2, it is conjectured that interleaving distance rather than frequency serves as the crucial avenue to promote classification performance. As expected, Experiment 3 documented that a new intervention containing personalized feedback along with brief instructions emphasizing the benefits of large interleaving over small interleaving significantly increased interleaving distance during self-regulated learning (d = 0.63) and improved classification accuracy accordingly (d = 0.51). Efforts to facilitate classification performance through the lens of interleaving distance could shed new light on the development of interventions in the field of inductive learning.

Insights into interleaving frequency: why learners switch to another category?

Interleaving frequency represents the decision to depart from a given category and move on to another. This raises the question: why does a learner temporarily terminate learning the prior category and focus on a new one? One possibility is that learners recognize that they will no longer gain further knowledge from the previous category if they persist in studying further exemplars from it. Metcalfe and Kornell (2005) argued that the decision to begin learning is guided by judgments of learning (JOLs), whereas the decision to stop depends on judgments of the rate of learning (JROLs). JOLs are prospective prediction of one's performance in future tests, while JROLs emphasize a dynamic process of information uptake. Regardless of the level of JOLs, when JROLs become relatively low, continuing to learn the same category is considered unproductive, thus learners tend to switch to another category.

Apart from the diminishing rate of perceived learning improvement, curiosity may also motivate learners to shift their focus away from the category they have just studied. As a motivational state, curiosity constantly stimulates leaners to seek new and unknown information (Gottlieb & Oudeyer, 2018;

Kidd & Hayden, 2015; Litman et al., 2005). Learners may drop a just-learned category because they are eager to explore the characteristics of other categories, especially at the early stages of learning. The results of our exploratory analyses speak to this curiosity-driven interleaving by showing that participants in both groups switched intensively between different categories at the outset of learning but did not maintain this high level of interleaving frequency thereafter (see the SM for details). Lu et al. (2021) documented a comparable trend by showing an initial high interleaving frequency followed by a rapid decline to stability. These observations imply that upon their initial encounter with a set of unfamiliar category labels, participants are naturally curious about the contents represented by each label, so they sequentially explore each one to reduce uncertainty. With a growing understanding of each category, participants' curiosity wanes, leading them to engage more in blocking to identify the commonalities within each category.

Insights into interleaving distance: leveraging informative switches for better category learning performance

With regard to interleaving distance, the current study encouraged participants to intermix exemplars to a greater extent, as interleaving distance significantly predicted classification accuracy in the post-intervention task (*r*s = .25-.31 across Experiments 1-3; see the SM). Large interleaving distances indeed offer distinct advantages and contribute importantly to inductive learning. Learners implemented a "fair" strategy by interleaving among all categories during their self-regulated learning, thereby gaining a preliminary impression of all to-be-studied categories (Kornell & Vaughn, 2018). However, the objective of employing large interleaving distances goes beyond merely providing a superficial overview of all categories. As learners sequentially view each category, they discern the structure of each category and how they relate to each other, for example, which categories are more similar, which are more readily

identifiable, and which are difficult to extract diagnostic features from. With a comprehensive understanding of all category structures, learners can then strategically allocate their limited learning opportunities to maximize learning efficiency.

While the current study confirmed that eliciting a larger interleaving distance was effective in enhancing classification performance, the framework of "desirable difficulties" suggests caution in its use. Defined by Bjork (1994), desirable difficulties entail creating learning conditions that, although challenging, ultimately foster more durable and flexible learning outcomes. Interleaved practice inherently represents a desirable difficulty. The interleaving distance further enables the modulation of learning difficulty by dynamically spacing exemplars of the same category apart, interspersed with a number of exemplars from different categories. Nonetheless, given the lag effect where temporal spacing between repetitions forms an inverted U-shaped relationship with final memory performance (Cepeda et al., 2009; Cepeda et al., 2008; Verkoeijen et al., 2005), a critical balance must be achieved for interleaving distance as well, as difficulties would become "undesirable" when they surpass the learner's capacity to benefit from them (Bjork & Bjork, 2011). For example, excessively long distances are likely to induce retrieval failure, resulting in poor inductive performance. Therefore, the effectiveness of large interleaving distances may be constrained by the forgetting of previously learned exemplars. Inspired by the progressive retrieval practice paradigm developed by Fiechter and Benjamin (2018, 2019), the desirable difficulties of study-phase retrieval can be methodically explored and determined by incrementally adding one unit of interleaving distance at a time.

Just like large interleaving distance, small interleaving distance also has its applications and limitations. Lu et al. (2021) demonstrated that interleaving was most beneficial when different categories shared similarities. Consistently, Abel (2023) found that frequent back and forth switches between highly

similar mushroom pairs, rather than among mushrooms with distinct appearances, predicted final classification accuracy. Switching between only two similar categories actually indicates a very low interleaving distance of 1, yet this kind of switching fully leverages the benefits of interleaving in facilitating discriminative contrast and enables swift detection of subtle differences between difficult-to-discriminate categories. Accordingly, small interleaving distance holds its unique value as long as the switches between categories are informative.

Nevertheless, it is crucial to emphasize that even when applied among similar categories, small interleaving distance may not be sufficiently informative, due to the omissions and biases it may cause. Imagine there are six categories labeled A-F, of which A, B and E are similar to each other and distinct from other categories, while the differences among C, D, and F are pronounced. A learner may initially observe that categories A and B are difficult to distinguish, leading to less informative alternation between these two categories because E was not even noticed. This behavior, in turn, prevents a direct comparison between A and B and another similar category, E. Put it differently, separating a small subset of categories from others (i.e., low interleaving distance) could narrow the learner's scope of attention, resulting in less ability to distinguish these two categories A and B from other similar categories.

Having discussed the different learning goals and impacts on learning associated with both large and small interleaving distances, it becomes apparent that strategically integrating and judiciously coordinating these methods may bring greater learning benefits. A practical approach to effectively managing various levels of interleaving entails that learners initially engage in broad interleaving during self-regulated learning. This initial phase helps them grasp the similarity structure, exemplar representativeness, and category concreteness across different categories. Subsequently, they can shift to narrower interleaving among similar categories to discern subtle distinctions, and also revisit categories

that are not readily retrievable due to forgetfulness. Finally, a return to broad interleaving allows for a comprehensive review of all categories, further reinforcing learning. This approach strikes a balance between global familiarization and thorough comparisons of categories. Given artists' paintings and butterfly species used in the current study are not particularly manipulated in terms of category structure, informative switches at short distances may have limited applicability and not affect the final classification performance. Because of this, we generally asked participants to increase their use of interleaving strategies and to separate repetitions of a given category by inserting as many exemplars from other categories as possible. However, for different experimental materials, future research should examine the boundary conditions of recommending large interleaving distance and develop more detailed intervention instructions targeting different stages of inductive learning.

Intervention effect on interleaving strategy transfer and its limitations

In addition to the elaboration on study sequence metrics outlined above, the current study also has implications for facilitating the transfer and generalization of self-regulated strategy use after the intervention. Transfer of learning strategies refers to the process through which a strategy acquired in the training context is applied or adapted to perform tasks or solve problems in a new or different context. Indeed, transfer of strategy use may occur implicitly, unfolding without any intentional effort to employ specific strategies acquired during training. For example, Chmielewski and Dansereau (1998) observed that students who received comprehensive and elaborate instructions on constructing and using knowledge maps improved their ability to process and recall textual information, even when they did not consciously report using the trained knowledge-mapping strategy in the transfer task. Considering the ultimate goal of teaching learning strategies is not just to enhance learning outcomes in the short term but to equip students with the skills necessary for subsequent learning and adaptation, understanding whether

and how students explicitly regulate learned strategies to new contexts is crucial for designing interventions that are truly beneficial in diverse learning environments.

A meta-analysis of instruction in study strategies revealed that teaching students about the effectiveness of some strategies, their methods of application, and the appropriate situations for their use significantly enhances academic performance in primary and secondary education (Donker et al., 2014). However, the absence of follow-up on students' use of strategies during self-regulated learning leaves it unclear whether the trained strategies were actually employed after the intervention. The current study bridges the gap between strategy instructions and subsequent performance by explicitly and objectively measuring the extent to which participants utilize the trained strategy in the post-intervention task. The results demonstrated explicit transfer: participants spontaneously interleaved more between different categories (Experiment 2) and juxtaposed a broader range of categories (Experiment 3) when transitioning from learning about artists' styles to butterfly species. Moreover, a mediation analysis for Experiment 3 indicated that the intervention effect on classification performance was completely mediated by its impact on the self-regulated usage of the interleaving strategy (represented by a composite indicator, which includes both interleaving frequency and interleaving distance), indirect effect = 0.04 [0.01, 0.08], $ab_{ps} = 0.29$; direct effect = 0.03 [-0.04, 0.10], $c'_{ps} = 0.20$ (see the SM). While the current study illuminates the specific process through which intervention instructions influence learning outcomes, it encounters limitations in two key areas.

Firstly, the instructional guidance in the current study explicitly informed participants that the interleaving strategy is superior to the blocking strategy and recommended participants use the interleaving strategy in the post-intervention task. When participants are explicitly instructed to employ a specific strategy in the target learning task, the transfer of this strategy frequently occurs with relative

ease (e.g., Chi et al., 1994; Wong et al., 2002). However, within the context of real-world learning and education, participants are likely to encounter situations that differ substantially from the target task, where the trained strategy might not prove to be as effective or easy to implement. Blindly employing it in such circumstances could potentially hinder the learning process. Essentially, any given strategy is advantageous under specific conditions. Regarding the interleaving strategy, Brunmair and Richter (2019) claimed that only confusable visual categories, similar to those used in the current study, take maximum advantage from it. Despite the widespread recognition of the interleaving effect, still a notable portion of learning materials fails to benefit from interleaved practice, including expository texts (Dobson, 2011), pronunciation rules (Carpenter & Mueller, 2013), and nuanced features shared by lists of objects (Sorensen & Woltz, 2016). As such, the generalizability of our intervention effects is somewhat constrained and warrants further investigation. To prevent inappropriate self-regulation of the strategy, furnishing learners with detailed and broad information regarding how, when, why, and where to use the strategy could help overcome this limitation (Borkowski et al., 1987; O'Sullivan & Pressley, 1984).

Finally, a fundamental issue is the longevity of newly-learned strategies. After a considerable amount of time has elapsed following an intervention, do learners remain capable of spontaneously applying the acquired strategies to new educationally relevant materials which are markedly different from those they encountered during training? Clearly, in the current study, the post-intervention task was administered immediately following the pre-intervention task and involved materials closely resembling those used in the prior task, that is, naturalistic visual categories. As a result, participants' enhanced application of the interleaving strategy during the post-intervention task can be deemed solely as near transfer. Even though some studies have attempted to extend the interval between the training and transfer tasks, these intervals have been largely limited to just a few days (e.g., Sun et al., 2022). Similarly, while efforts have been

undertaken to incorporate materials of an educational nature in transfer tasks, they substantially lack the complexity of assignments that students are actually expected to complete (e.g., Chmielewski & Dansereau, 1998; Cook & Mayer, 1988). Therefore, an important gap remains before strategy training can be universally recommended for use in educational settings.

Concluding Remarks

An instructional intervention consisting of individualized feedback on blocked vs. interleaved performance and a summary of research findings regarding the interleaving effect is effective in correcting the metacognitive illusion that blocking is superior to interleaving, enhancing self-employment of interleaving, and increasing the percentage of learners who demonstrate an inclination to interleave category exemplars. However, this intervention is insufficient to elicit a reliable increase in interleaving distance and produces minimal benefit to classification performance. Providing individual feedback and adding a summary of research findings relating to the benefits of large interleaving (over small interleaving and pure blocking) successfully boosts interleaving distance and yields a medium-sized potentiating effect on classification accuracy. The intervention effects on metacognitive awareness, strategy utilization, interleaving distance, and inductive performance transfer to category learning in new domains, highlighting the potential for these intervention instructions to be applied across a variety of disciplines. These findings motivate educators and curriculum designers to explicitly integrate the benefits of interleaving, particularly extensive interleaving, into teaching practices to facilitate inductive learning.

References

Abel, R. (2023). Some fungi are not edible more than once: The impact of motivation to avoid confusion on learners' study sequence choices. *Journal of Applied Research in Memory and Cognition*. https://doi.org/10.1037/mac0000107

Ariel, R., Dunlosky, J., & Bailey, H. (2009). Agenda-based regulation of study-time allocation: when agendas override item-based monitoring. *Journal of Experimental Psychology: General*, *138*(3), 432-447. https://doi.org/10.1037/a0015928

Badali, S., Rawson, K. A., & Dunlosky, J. (2022). Do Students Effectively Regulate Their Use of Self-Testing as a Function of Item Difficulty? *Educational Psychology Review*, *34*(3), 1651-1677. https://doi.org/10.1007/s10648-022-09665-6

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

Bjork, E. L., & Bjork, R. A. (2011). Making things hard on yourself, but in a good way: Creating desirable difficulties to enhance learning. In M. A. Gernsbacher, R. W. Pew, L. M. Hough, & J. R. Pomerantz (Eds.), *Psychology and the real world: Essays illustrating fundamental contributions to society* (pp. 59-68). Worth Publishers.

Bjork, R. A. (1994). Memory and metamemory considerations in the training of human beings. In J. Metcalfe & A. Shimamura (Eds.), *Metacognition: Knowing about knowing* (pp. 185–205), MIT Press.

Bjork, R. A., Dunlosky, J., & Kornell, N. (2013). Self-Regulated Learning: Beliefs, Techniques, and Illusions. *Annual Review of Psychology*, *64*(1), 417-444. https://doi.org/10.1146/annurev-psych-

113011-143823

Borkowski, J. G., Carr, M., & Pressley, M. (1987). “Spontaneous” strategy use: Perspectives from metacognitive theory. *Intelligence*, *11*(1), 61-75. https://doi.org/https://doi.org/10.1016/0160-2896(87)90027-4

Brunmair, M., & Richter, T. (2019). Similarity matters: A meta-analysis of interleaved learning and its moderators. *Psychological Bulletin*, *145*(11), 1029-1052. https://doi.org/10.1037/bul0000209

Carpenter, S. K., & Mueller, F. E. (2013). The effects of interleaving versus blocking on foreign language pronunciation learning. *Memory & Cognition*, *41*(5), 671-682. https://doi.org/10.3758/s13421-012-0291-4

Carvalho, P. F., & Goldstone, R. L. (2014a). Effects of interleaved and blocked study on delayed test of category learning generalization. *Frontiers in Psychology*, *5*, 936. https://doi.org/10.3389/fpsyg.2014.00936

Carvalho, P. F., & Goldstone, R. L. (2014b). Putting category learning in order: Category structure and temporal arrangement affect the benefit of interleaved over blocked study. *Memory & Cognition*, *42*(3), 481-495. https://doi.org/10.3758/s13421-013-0371-0

Cepeda, N. J., Coburn, N., Rohrer, D., Wixted, J. T., Mozer, M. C., & Pashler, H. (2009). Optimizing distributed practice: theoretical analysis and practical implications. *Experimental Psychology*, *56*(4), 236-246. https://doi.org/10.1027/1618-3169.56.4.236

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. https://doi.org/10.1037/0033-2909.132.3.354

Cepeda, N. J., Vul, E., Rohrer, D., Wixted, J. T., & Pashler, H. (2008). Spacing Effects in Learning: A

Temporal Ridgeline of Optimal Retention. *Psychological Science*, *19*(11), 1095-1102. https://doi.org/10.1111/j.1467-9280.2008.02209.x

Chi, M. T. H., De Leeuw, N., Chiu, M.-H., & Lavancher, C. (1994). Eliciting self-explanations improves understanding. *Cognitive Science*, *18*(3), 439-477. https://doi.org/https://doi.org/10.1016/0364-0213(94)90016-7

Chmielewski, T. C., & Dansereau, D. F. (1998). Enhancing the recall of text: Knowledge mapping training promotes implicit transfer. *Journal of Educational Psychology*, *90*(3), 407-413. https://doi.org/https://doi.org/10.1037/0022-0663.90.3.407

Cook, L. K., & Mayer, R. E. (1988). Teaching Readers About the Structure of Scientific Text. *Journal of Educational Psychology*, *80*(4), 448-456. https://doi.org/10.1037/0022-0663.80.4.448

Dobson, J. L. (2011). Effect of selected "desirable difficulty" learning strategies on the retention of physiology information. *Advances in Physiology Education*, *35*(4), 378-383. https://doi.org/10.1152/advan.00039.2011

Donker, A. S., de Boer, H., Kostons, D., Dignath van Ewijk, C. C., & van der Werf, M. P. C. (2014). Effectiveness of learning strategy instruction on academic performance: A meta-analysis. *Educational Research Review*, *11*, 1-26. https://doi.org/10.1016/j.edurev.2013.11.002

Dunlosky, J., & Hertzog, C. (2000). Updating knowledge about encoding strategies: A componential analysis of learning about strategy effectiveness from task experience. *Psychology and Aging*, *15*(3), 462-474. https://doi.org/https://doi.org/10.1037/0882-7974.15.3.462

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

Dunlosky, J., & Thiede, K. W. (2004). Causes and constraints of the shift-to-easier-materials effect in the control of study. *Memory & Cognition*, *32*(5), 779-788. https://doi.org/10.3758/BF03195868

Eglington, L. G., & Kang, S. H. K. (2017). Interleaved Presentation Benefits Science Category Learning. *Journal of Applied Research in Memory and Cognition*, *6*(4), 475-485. https://doi.org/10.1016/j.jarmac.2017.07.005

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

Fiechter, J. L., & Benjamin, A. S. (2018). Diminishing-cues retrieval practice: A memory-enhancing technique that works when regular testing doesn't. *Psychonomic Bulletin & Review*, *25*(5), 1868-1876. https://doi.org/10.3758/s13423-017-1366-9

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

Firth, J., Rivers, I., & Boyle, J. (2021). A systematic review of interleaving as a concept learning strategy. *Review of Education*, *9*(2), 642-684. https://doi.org/10.1002/rev3.3266

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. https://doi.org/10.3758/s13421-019-00918-4

Gluckman, M., Vlach, H. A., & Sandhofer, C. M. (2014). Spacing Simultaneously Promotes Multiple Forms of Learning in Children's Science Curriculum. *Applied Cognitive Psychology*, *28*(2), 266-273.

https://doi.org/10.1002/acp.2997

Gottlieb, J., & Oudeyer, P. Y. (2018). Towards a neuroscience of active sampling and curiosity. *Nature Reviews Neuroscience*, *19*(12), 758-770. https://doi.org/10.1038/s41583-018-0078-0

Guzman-Munoz, F. J. (2017). The advantage of mixing examples in inductive learning: a comparison of three hypotheses. *Educational Psychology*, *37*(4), 421-437. https://doi.org/10.1080/01443410.2015.1127331

Holland, J. H., Holyoak, K. J., Nisbett, R. E., & Thagard, P. R. (1989). *Induction: Processes of inference, learning, and discovery*. MIT press.

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

Janssen, E. M., van Gog, T., van de Groep, L., de Lange, A. J., Knopper, R. L., Onan, E., Wiradhany, W., & de Bruin, A. B. H. (2023). The Role of Mental Effort in Students' Perceptions of the Effectiveness of Interleaved and Blocked Study Strategies and Their Willingness to Use Them. *Educational Psychology Review*, *35*(3), 85. https://doi.org/10.1007/s10648-023-09797-3

JASP Team. (2022). JASP (Version 0.16.2)[Computer software]. https://jasp-stats.org/

Jeffreys, H. (1961). *The theory of probability*. Oxford University Press.

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. https://doi.org/10.1002/acp.1801

Karpicke, J. D. (2009). Metacognitive control and strategy selection: Deciding to practice retrieval during learning. *Journal of Experimental Psychology: General*, *138*(4), 469-486.

https://doi.org/10.1037/a0017341

Kass, R. E., & Raftery, A. E. (1995). Bayes Factors. *Journal of the American Statistical Association*, *90*(430), 773-795. https://doi.org/10.2307/2291091

Kensinger, E. A., & Corkin, S. (2003). Effect of negative emotional content on working memory and long-term memory. *Emotion*, *3*(4), 378-393. https://doi.org/10.1037/1528-3542.3.4.378

Kidd, C., & Hayden, B. Y. (2015). The Psychology and Neuroscience of Curiosity. *Neuron*, *88*(3), 449-460. https://doi.org/10.1016/j.neuron.2015.09.010

Koriat, A., & Bjork, R. A. (2005). Illusions of competence in monitoring one's knowledge during study. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31*(2), 187-194. https://doi.org/10.1037/0278-7393.31.2.187

Koriat, A., & Bjork, R. A. (2006). Illusions of competence during study can be remedied by manipulations that enhance learners' sensitivity to retrieval conditions at test. *Memory & Cognition*, *34*(5), 959-972. https://doi.org/10.3758/BF03193244

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

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. https://doi.org/10.1037/a0017807

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. https://doi.org/10.1037/arc0000042

Litman, J. A., Hutchins, T. L., & Russon, R. K. (2005). Epistemic curiosity, feeling-of-knowing, and exploratory behaviour. *Cognition & Emotion*, *19*(4), 559-582. https://doi.org/10.1080/02699930441000427

Lu, X., Penney, T. B., & Kang, S. H. K. (2021). Category similarity affects study choices in self-regulated learning. *Memory & Cognition*, *49*(1), 67-82. https://doi.org/10.3758/s13421-020-01074-w

McDaniel, M. A., & Einstein, G. O. (2020). Training Learning Strategies to Promote Self-Regulation and Transfer: The Knowledge, Belief, Commitment, and Planning Framework. *Perspectives on Psychological Science*, *15*(6), 1363-1381. https://doi.org/10.1177/1745691620920723

Metcalfe, J., & Kornell, N. (2003). The dynamics of learning and allocation of study time to a region of proximal learning. *Journal of Experimental Psychology: General*, *132*(4), 530-542. https://doi.org/10.1037/0096-3445.132.4.530

Metcalfe, J., & Kornell, N. (2005). A Region of Proximal Learning model of study time allocation. *Journal of Memory and Language*, *52*(4), 463-477. https://doi.org/10.1016/j.jml.2004.12.001

Middlebrooks, C. D., & Castel, A. D. (2018). Self-regulated learning of important information under sequential and simultaneous encoding conditions. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *44*(5), 779-792. https://doi.org/10.1037/xlm0000480

Mielicki, M. K., & Wiley, J. (2022). Exploring the necessary conditions for observing interleaved practice benefits in math learning. *Learning and Instruction*, *80*. https://doi.org/10.1016/j.learninstruc.2022.101583

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

Nakata, T., & Suzuki, Y. (2019). Mixing Grammar Exercises Facilitates Long-Term Retention: Effects of Blocking, Interleaving, and Increasing Practice. *The Modern Language Journal*. https://doi.org/10.1111/modl.12581

O'Sullivan, J. T., & Pressley, M. (1984). Completeness of instruction and strategy transfer. *Journal of Experimental Child Psychology*, *38*(2), 275-288. https://doi.org/https://doi.org/10.1016/0022-0965(84)90126-7

Onan, E., Wiradhany, W., Biwer, F., Janssen, E. M., & de Bruin, A. B. H. (2022). Growing Out of the Experience: How Subjective Experiences of Effort and Learning Influence the Use of Interleaved Practice. *Educational Psychology Review*, *34*(4), 2451-2484. https://doi.org/10.1007/s10648-022-09692-3

Pan, S. C., Tajran, J., Lovelett, J., Osuna, J., & Rickard, T. C. (2019). Does interleaved practice enhance foreign language learning? The effects of training schedule on Spanish verb conjugation skills. *Journal of Educational Psychology*, *111*(7), 1172-1188. https://doi.org/10.1037/edu0000336

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

Rawson, K. A., Thomas, R. C., & Jacoby, L. L. (2015). The Power of Examples: Illustrative Examples Enhance Conceptual Learning of Declarative Concepts. *Educational Psychology Review*, *27*(3), 483-504. https://doi.org/10.1007/s10648-014-9273-3

Rohrer, D., Dedrick, R. F., & Burgess, K. (2014). The benefit of interleaved mathematics practice is not limited to superficially similar kinds of problems. *Psychonomic Bulletin & Review*, *21*(5), 1323-1330. https://doi.org/10.3758/s13423-014-0588-3

Rohrer, D., Dedrick, R. F., & Stershic, S. (2015). Interleaved practice improves mathematics learning. *Journal of Educational Psychology*, *107*(3), 900-908. https://doi.org/10.1037/edu0000001

Rohrer, D., & Taylor, K. (2007). The shuffling of mathematics problems improves learning. *Instructional Science*, *35*(6), 481-498. https://doi.org/10.1007/s11251-007-9015-8

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. https://doi.org/10.1037/edu0000119

Shanks, D. R., Don, H. J., Boustani, S., & Yang, C. (2023). Test-Enhanced Learning. In *Oxford Research Encyclopedia of Psychology*. https://doi.org/10.1093/acrefore/9780190236557.013.908

Sorensen, L. J., & Woltz, D. J. (2016). Blocking as a friend of induction in verbal category learning. *Memory & Cognition*, *44*(7), 1000-1013. https://doi.org/10.3758/s13421-016-0615-x

Sun, Y., Shi, A., Zhao, W., Yang, Y., Li, B., Hu, X., Shanks, D. R., Yang, C., & Luo, L. (2022). Long-Lasting Effects of an Instructional Intervention on Interleaving Preference in Inductive Learning and Transfer. *Educational Psychology Review*, *34*(3), 1679-1707. https://doi.org/10.1007/s10648-022-09666-5

Suzuki, Y., Yokosawa, S., & Aline, D. (2020). The role of working memory in blocked and interleaved grammar practice: Proceduralization of L2 syntax. *Language Teaching Research*, *26*(4), 671-695. https://doi.org/10.1177/1362168820913985

Tauber, S. K., Dunlosky, J., Rawson, K. 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*(2), 356-363. https://doi.org/10.3758/s13423-012-0319-6

Taylor, K., & Rohrer, D. (2010). The effects of interleaved practice. *Applied Cognitive Psychology*, *24*(6), 837-848. https://doi.org/10.1002/acp.1598

Verkoeijen, P. P., & Bouwmeester, S. (2014). Is spacing really the "friend of induction"? *Frontiers in Psychology*, *5*, 259. https://doi.org/10.3389/fpsyg.2014.00259

Verkoeijen, P. P., Rikers, R. M., & Schmidt, H. G. (2005). Limitations to the spacing effect: demonstration of an inverted u-shaped relationship between interrepetition spacing and free recall. *Experimental Psychology*, *52*(4), 257-263. https://doi.org/10.1027/1618-3169.52.4.257

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. https://doi.org/10.3758/s13421-010-0063-y

Wong, R. M. F., Lawson, M. J., & Keeves, J. (2002). The effects of self-explanation training on students' problem solving in high-school mathematics. *Learning and Instruction*, *12*(2), 233-262. https://doi.org/https://doi.org/10.1016/S0959-4752(01)00027-5

Wong, S. S. H., Chen, S., & Lim, S. W. H. (2021). Learning melodic musical intervals: To block or to interleave? *Psychology of Music*, *49*(4), 1027-1046. https://doi.org/10.1177/0305735620922595

Wong, S. S. H., Low, A. C. M., Kang, S. H. K., & Lim, S. W. H. (2020). Learning Music Composers' Styles: To Block or to Interleave? *Journal of Research in Music Education*, *68*(2), 156-174. https://doi.org/10.1177/0022429420908312

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. https://doi.org/10.1037/xge0000177

Yan, V. X., & Sana, F. (2019). Interleaving Benefits the Learning of Complex Perceptual Categories: Evidence Against the Discriminative-Contrast Hypothesis. *Journal of Cognitive Education and Psychology*, *18*(1), 35-51. https://doi.org/10.1891/1945-8959.18.1.35

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. https://doi.org/10.1037/xap0000139

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

Yang, C., Potts, R., & Shanks, D. R. (2017). Metacognitive unawareness of the errorful generation benefit and its effects on self-regulated learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *43*(7), 1073-1092. https://doi.org/10.1037/xlm0000363

Yang, C., & Shanks, D. R. (2018). The forward testing effect: Interim testing enhances inductive learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *44*(3), 485-492. https://doi.org/10.1037/xlm0000449

Yang, C., Sun, B., & Shanks, D. R. (2018). The anchoring effect in metamemory monitoring. *Memory & Cognition*, *46*(3), 384-397. https://doi.org/10.3758/s13421-017-0772-6

Zulkiply, N. (2013). Effect of Interleaving Exemplars Presented as Auditory Text on Long- term Retention in Inductive Learning. *Procedia - Social and Behavioral Sciences*, *97*, 238-245. https://doi.org/10.1016/j.sbspro.2013.10.228

Zulkiply, N., & Burt, J. S. (2013a). The exemplar interleaving effect in inductive learning: moderation by the difficulty of category discriminations. *Memory & Cognition*, *41*(1), 16-27. https://doi.org/10.3758/s13421-012-0238-9

Zulkiply, N., & Burt, J. S. (2013b). Inductive Learning: Does Interleaving Exemplars Affect Long-Term Retention? *Malaysian Journal of Learning and Instruction*, *10*, 133-155.

https://doi.org/10.32890/mjli.10.2013.7655

Zulkiply, N., McLean, J., Burt, J. S., & Bath, D. (2012). Spacing and induction: Application to exemplars presented as auditory and visual text. *Learning and Instruction*, *22*(3), 215-221. https://doi.org/10.1016/j.learninstruc.2011.11.002

Table 1. Metacognitive judgements of the relative effectiveness of study strategies in Experiments 1-3

Experiment	Judgement choice	Before intervention	After Intervention	Test of difference between choices made before and after the intervention phase	*p*
Experiment 1					
Control	B > I	55%	55%		1.000[a]
	B = I	13%	13%		
	B < I	32%	32%		
Intervention	B > I	59%	28%		.018[a]
	B = I	9%	6%		
	B < I	31%	66%		
Experiment 2					
Control	B > I	62%	56%		.937[a]
	B = I	12%	12%		
	B < I	27%	32%		
Intervention	B > I	56%	18%		.002[a]
	B = I	6%	6%		
	B < I	38%	77%		
Experiment 3					
Control	B	64%	46%	3.25	.197
	SI	15%	33%		
	LI	21%	21%		
Intervention	B	55%	12%	20.14	< .001
	SI	30%	21%		
	LI	15%	67%		

Note: In Experiments 1 and 2, B > I refers to the belief that blocking is superior to interleaving; B = I refers to the belief that the two strategies are comparable in efficacy; B < I refers to the belief that

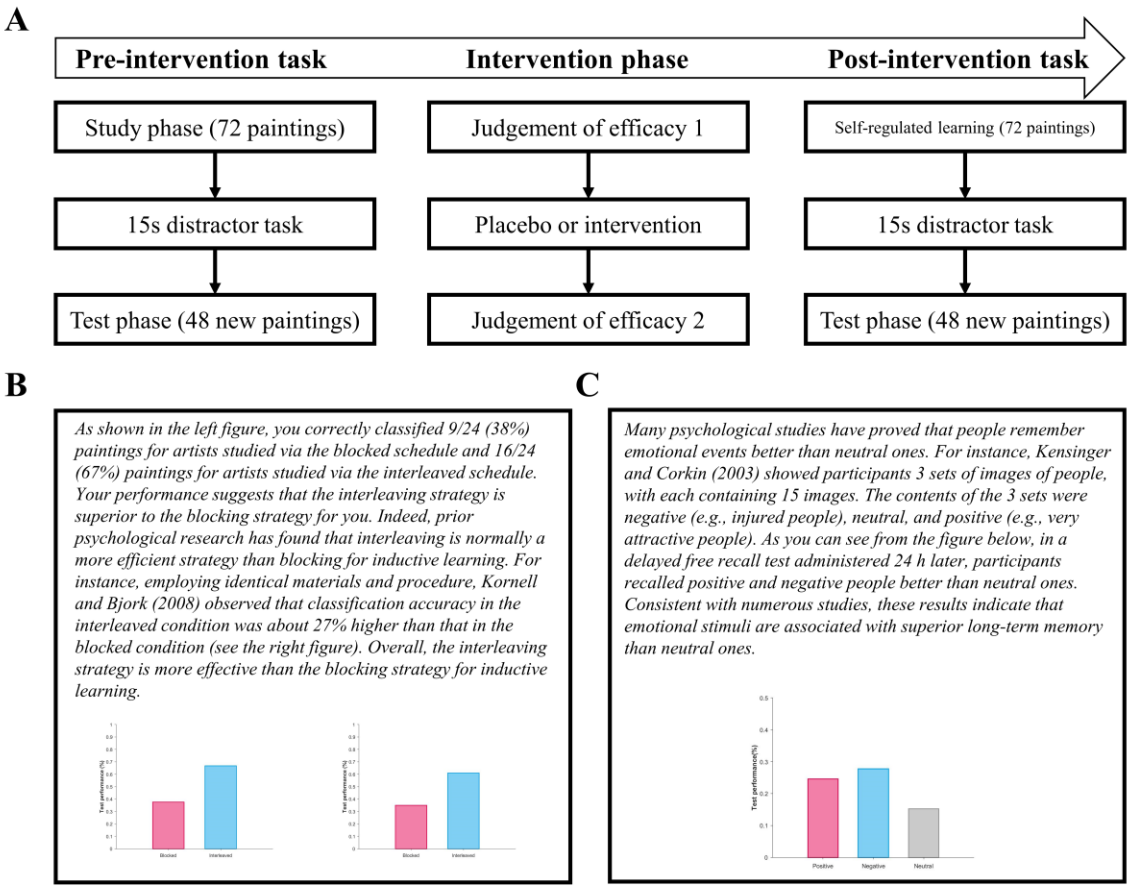
blocking is inferior to interleaving. In Experiment 3, B represents the belief that blocking is the most effective strategy; SI represents the belief that small interleaving is the most effective strategy; LI represents the belief that large interleaving is the most effective strategy. Chi-squared tests or Fisher-Freeman-Halton's exact tests (if the prerequisites for a chi-squared test were not met due to the predicted frequency in some cells being too small) were conducted to examine the difference between metacognitive judgements made before and after the intervention phase in both the control and intervention groups. *p*-values marked with a superscript "a" indicate that the Fisher-Freeman-Halton's exact test was conducted. This test only generates a *p*-value and not any associated test statistic.

Table 2. Interleaving preference and classification performance in the post-intervention task of Experiments 1-3

Experiment	Measure	Control	Intervention	Difference & 95% CI	t/χ^2	*df*	*p*	*d*
Experiment 1	Interleaving frequency	39% (31%)	56% (33%)	17% [0.5%, 33%]	2.06	61	.044	0.52
	Percentage of interleavers	29%	56%	27%	4.76	1	.029	-
	Interleaving distance (out of 11)	5.41(3.49)	6.26 (3.64)	0.84 [-0.96, 2.64]	0.94	61	.353	0.24
	Test performance	68% (22%)	72% (24%)	4% [-8%, 15%]	0.61	61	.547	0.18
Experiment 2	Interleaving frequency	42% (29%)	64% (33%)	22% [7%, 37%]	2.97	66	.004	0.72
	Percentage of interleavers	32%	71%	38%	9.95	1	.002	-
	Interleaving distance (out of 15)	7.48 (4.44)	8.81 (4.79)	1.33 [-0.91, 3.57]	1.19	66	.239	0.29
	Test performance	41% (14%)	39% (14%)	-2% [-9%, 5%]	-0.49	66	.629	-0.14
Experiment 3	Interleaving frequency	40% (25%)	68% (33%)	27% [13%, 42%]	3.80	64	<.001	0.94
	Percentage of interleavers	39%	70%	30%	6.11	1	.013	-
	Interleaving distance (out of 15)	7.22 (4.78)	9.89 (3.61)	2.67 [0.59, 4.75]	2.57	59.53	.013	0.63
	Test performance	50% (16%)	57% (11%)	7% [0.3%, 14%]	2.07	64	.043	0.51

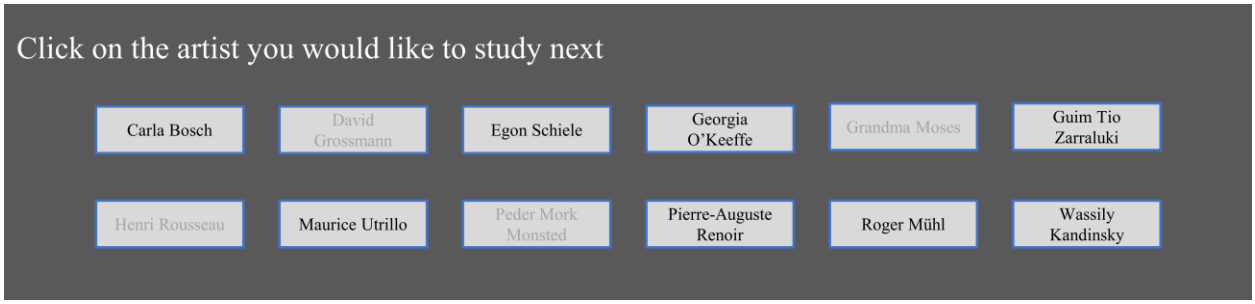
Note: *SD*s are shown in parentheses. For interleaving frequency, interleaving distance, and test performance, independent *t*-tests were conducted to compare the difference between the intervention and control groups. For percentage of interleavers, chi-squared tests were performed to compare the difference between the two groups.

Figure 1. Procedure and instructional intervention in Experiment 1



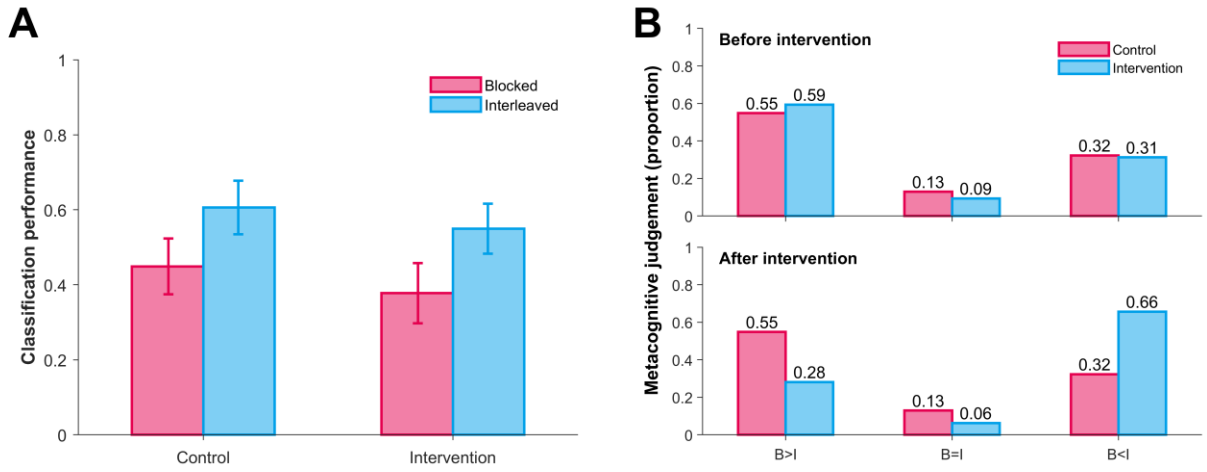
Note: A: Schematic diagram of the procedure in Experiment 1. B: A sample of the intervention instructions and figures viewed by participants in the intervention group. C: Control instructions and figure viewed by participants in the control group.

Figure 2. Selection interface in the self-regulated study phase of the post-intervention task in Experiment 1



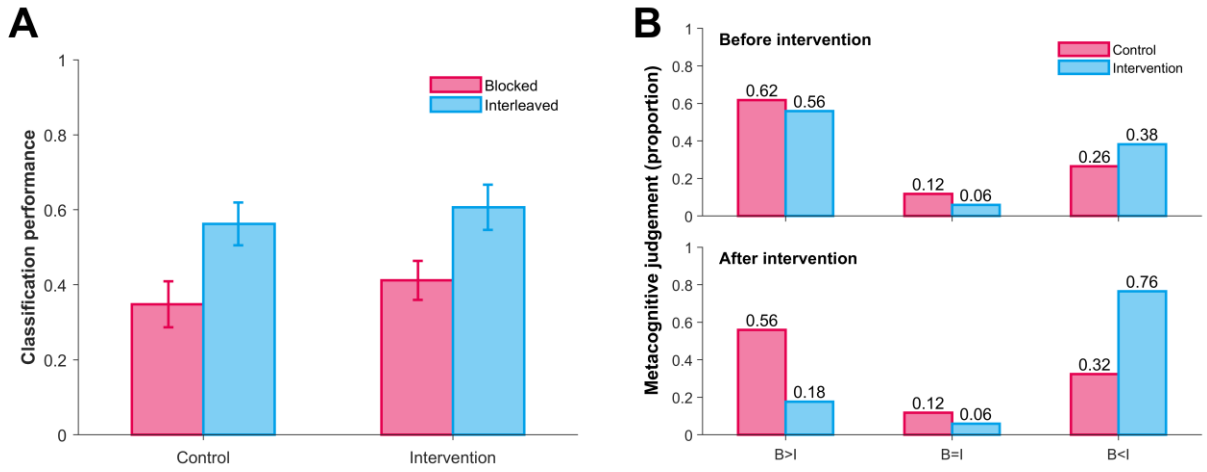
Note: Twelve artists' names are aligned in two rows at the center of the screen, waiting for participants to make their next choice. Specifically, names in black indicate that there are paintings left to study from these artists, whereas names in grey indicate there are no more paintings left to study from these artists. In the experiment, all artists' names were replaced by popular Chinese boy names.

Figure 3. Classification performance in the pre-intervention task and metacognitive judgments in Experiment 1



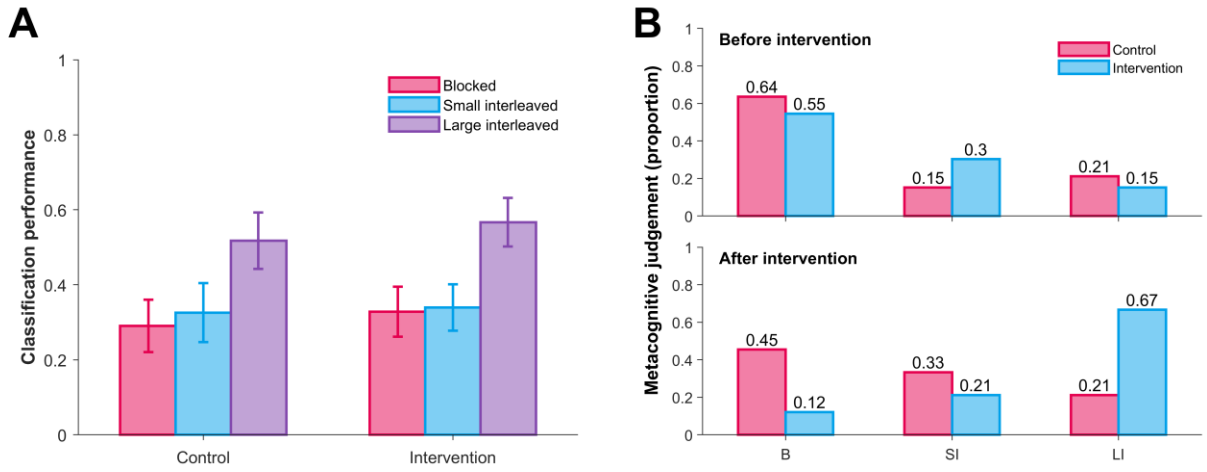
Note: A: Classification performance in the pre-intervention task as a function of group and study schedule. Error bars represent 95% CI. B: Metacognitive judgments made before and after the intervention phase on study strategy effectiveness in the control and intervention groups. In Panel B, the upper section illustrates metacognitive judgments made prior to the intervention phase, while the lower section depicts judgments made after the intervention phase. The y-axis shows the proportion of participants believing that blocking is superior to interleaving (B > I), the proportion believing that blocking is approximately equal to interleaving (B = I), and the proportion believing that interleaving is superior to blocking (B < I).

Figure 4. Classification performance in the pre-intervention task and metacognitive judgments in Experiment 2



Note: A: Classification performance in the pre-intervention task as a function of group and study schedule. Error bars represent 95% CI. B: Metacognitive judgments made before and after the intervention phase on study strategy effectiveness in the control and intervention groups. In Panel B, the upper section illustrates metacognitive judgments made prior to the intervention phase, while the lower section depicts judgments made after the intervention phase. The y-axis shows the proportion of participants believing that blocking is superior to interleaving (B > I), the proportion believing that blocking is approximately equal to interleaving (B = I), and the proportion believing that interleaving is superior to blocking (B < I).

Figure 5. Classification performance in the pre-intervention task and metacognitive judgments in Experiment 3



Note: A: Classification performance in the pre-intervention task as a function of group and study schedule. Error bars represent 95% CI. B: Metacognitive judgments made before and after the intervention phase on study strategy effectiveness in control and intervention groups. In Panel B, the upper section illustrates metacognitive judgments made prior to the intervention phase, while the lower section depicts judgments made after the intervention phase. The y-axis shows the proportion of participants believing that blocking (B) is the optimal strategy, the proportion believing that small interleaving (SI) is the optimal strategy, and the proportion believing that large interleaving (LI) is the optimal strategy.

Supplemental Materials

Experiment 1

Exact wording of the instructions

The intervention instructions were adapted according to each participant's test performance in the pre-intervention task as follows:

For participants who correctly classified more paintings in the interleaved than in the blocked condition, the instructions were as follows: *As shown in the left figure, you correctly classified X/24 (X%) paintings for artists studied via the blocked schedule and X/24 (X%) paintings for artists studied via the interleaved schedule. Your performance suggests that the interleaving strategy is superior to the blocking strategy for you. Indeed, prior psychological research has found that interleaving is normally a more efficient strategy than blocking for inductive learning. For instance, employing identical materials and procedure, Kornell and Bjork (2008) observed that classification accuracy in the interleaved condition was about 27% higher than that in the blocked condition (see the right figure). Overall, the interleaving strategy is more effective than the blocking strategy for inductive learning.*

For participants who correctly classified more paintings in the blocked than in the interleaved condition, the instructions were as follows: *As shown in the left figure, you correctly classified X/24 (X%) paintings for artists studied via the blocked schedule and X/24 (X%) paintings for artists studied via the interleaved schedule. Your performance suggests that the blocking strategy seems to be superior to the interleaving strategy for you. However, prior psychological research has found that interleaving is normally a more efficient strategy than blocking for inductive learning. For instance, employing identical materials and procedure, Kornell and Bjork (2008) observed that classification accuracy in the interleaved condition was about 27% higher than that in the blocked condition (see the right figure).*

Although your performance is not consistent with previous findings, this might just be a result of random variation. Overall, numerous prior studies consistently demonstrated that the interleaving strategy is more effective than the blocking strategy for inductive learning.

For participants who correctly classified an equal number of paintings in the interleaved and blocked conditions, the instructions were as follows: *As shown in the left figure, you correctly classified X/24 (X%) paintings for artists studied via the blocked schedule and X/24 (X%) paintings for artists studied via the interleaved schedule. Your performance suggests that the blocking and interleaving strategies seem to be equally effective for you. However, prior psychological research has found that interleaving is normally a more efficient strategy than blocking for inductive learning. For instance, employing identical materials and procedure, Kornell and Bjork (2008) observed that classification accuracy in the interleaved condition was about 27% higher than that in the blocked condition (see the right figure). Although your performance is not consistent with previous findings, this might just be a result of random variation. Overall, numerous prior studies consistently demonstrated that the interleaving strategy is more effective than the blocking strategy.*

The placebo instructions were as follows: *Next, please read a short passage pertaining to emotion and memory. Many psychological studies have proved that people remember emotional events better than neutral ones. For instance, Kensinger and Corkin (2003) showed participants 3 sets of images of people, with each containing 15 images. The contents of the 3 sets were negative (e.g., injured people), neutral, and positive (e.g., very attractive people). As you can see from the figure below, in a delayed free recall test administered 24 h later, participants recalled positive and negative people better than neutral ones. Consistent with numerous studies, these results indicate that emotional stimuli are associated with superior long-term memory than neutral ones.*

Exploratory analyses

Interleaving frequency as a function of time period

After excluding 11 compulsory choices from 71 choices, each participant could choose the to-be-learned category actively 60 times in the post-intervention task. These 60 choices were divided into 6 periods, with the first period comprising the first 10 choices, the second period comprising the second 10 choices, and so on. The proportion of switches was calculated for each period in both the control and intervention groups. Figure S1 shows how employment of interleaving changes over time in the two groups.

A two-way mixed-factor ANOVA was performed with group as the between-subjects factor and period as the within-subjects factor (Greenhouse–Geisser corrections were implemented if the assumption of sphericity was violated). There was a main effect of group, $F(1, 61) = 4.24$, $p = .044$, $\eta_p^2 = 0.065$, with participants in the intervention group interleaving more frequently than those in the control group. There was also a main effect of period, $F(3.09, 188.29) = 3.56$, $p = .014$, $\eta_p^2 = 0.055$, suggesting that participants switched more frequently at the beginning of the study phase and transitioned from interleaving to blocking across the study phase. This strategy shift is similar to that reported by Lu et al. (2021). The interaction term was not statistically significant, $F(3.09, 188.29) = 1.15$, $p = .329$, $\eta_p^2 = .019$, suggesting that the intervention effect on interleaving preference was maintained across a prolonged study phase.

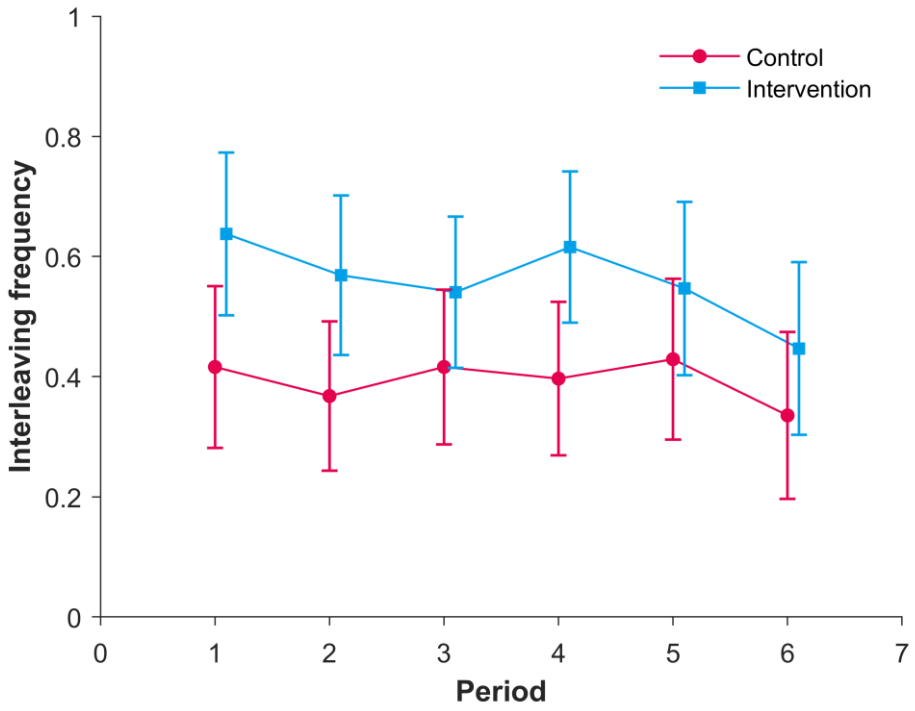


Figure S1. Interleaving frequency in the two groups as a function of time period in Experiment 1. Error bars represent 95% CI.

Interleaving frequency and distance as predictors of classification performance

Despite the absence of intervention effects on interleaving distance and classification accuracy in the post-intervention task, we posited that both interleaving frequency and interleaving distance could serve as potential predictors of classification performance. Due to the substantial overlap between these study sequence metrics, their correlations with classification performance were analyzed separately. Table S1 presents the correlations among interleaving frequency, interleaving distance, and classification performance in the post-intervention task for both the control and intervention groups in each experiment.

In the control group, both interleaving frequency and interleaving distance positively correlated with classification accuracy in the post-intervention task. In the intervention group, positive correlations were also observed for interleaving frequency and interleaving distance though they were not statistically significant. Because the result patterns were similar across the two groups, we collapsed the data to increase statistical power. In this analysis, the correlation between interleaving frequency and classification accuracy reached significance and the same was true for the correlation between interleaving distance and classification accuracy. Given that both interleaving frequency and distance are positive predictors of classification performance across the two groups, the intervention targeting to enhance classification performance by stimulating frequent switches or increasing interleaving distance should be effective.

Table S1. Correlations between study sequence metrics and classification accuracy in the post-intervention task for Experiments 1-3

Experiment	*r*(Frq, Dst)	*r*(Frq, Acc)	*r*(Dst, Acc)
Experiment 1			
Control	.85***	.36*	.43*
Intervention	.77***	.21	.21
All groups	.81***	.29*	.31*
Experiment 2			
Control	.84***	.21	.35*
Intervention	.77***	.31	.18
All groups	.79***	.23	.25*
Experiment 3			
Control	.74***	.42*	.15
Intervention	.63***	.38*	.45**
All groups	.70***	.44***	.31**

Note: Pairwise correlations between interleaving frequency (Frs), interleaving distance (Dst), and classification accuracy (Acc) in the post-intervention task were calculated separately for the control and intervention groups. * $p < .05$; ** $p < .01$; *** $p < .001$.

Interleaving effect and magnitude of the instructional intervention effect

We explored whether exhibiting an interleaving effect in the pre-intervention task affected the magnitude of the instructional intervention effect in Experiment 1. Participants in the intervention group were split into two subgroups according to whether their test performance was better in the interleaved than in the blocked condition in the induction test of the pre-intervention task. Although statistical results were not reliable since only a few participants (five out of 32) manifested a reversed interleaving effect, it can nevertheless be seen from the descriptive data in Table S2 that these individuals were less likely to correct their metacognitive beliefs and continued to use the blocked strategy in the post-intervention task than those who did demonstrate an interleaving effect in the pre-intervention task. This is consistent with the idea that participants' beliefs about their individual uniqueness precluded them from altering their preference (Yan et al., 2016).

Table S2. Metacognitive judgments and strategy employment in subgroups split by test performance in the pre-intervention task in Experiments 1 and 2

Experiment	Test performance in the pre-intervention task	Metacognitive judgements	Before intervention	After intervention	Interleaving frequency	Percentage of interleavers	Interleaving distance
Experiment 1							
5 participants	B ≥ I	B > I	40%	60%	55% (31%)	60%	6.35 (3.55)
		B = I	0	0			
		B < I	60%	40%			
27 participants	B < I	B > I	63%	22%	56% (33%)	56%	7.42 (3.64)
		B = I	11%	7%			
		B < I	26%	70%			
Experiment 2							
4 participants	B ≥ I	B > I	100%	75%	42% (34%)	50%	6.26 (4.72)
		B = I	0	0			
		B < I	0	25%			
30 participants	B < I	B > I	50%	10%	67% (33%)	73%	10.29 (4.79)
		B = I	7%	7%			
		B < I	43%	83%			

Note: *SD*s are shown in parentheses. All participants in the table are from the intervention group. In the second column, B ≥ I refers to participants who did not show an interleaving effect in the test stage of the pre-intervention task and B < I refers to those who demonstrated an interleaving effect.

Experiment 2

Exploratory analyses

Interleaving frequency as a function of time period

The total number of 48 discretionary choices in the post-intervention task was divided into 6 periods, each including 8 choices. A mixed-factor ANOVA on the proportion of switches (i.e., interleaving frequency) was performed with Greenhouse–Geisser corrections. The main effect of group was significant, $F(1, 66) = 8.81$, $p = .004$, $\eta_p^2 = .118$. The effect of time period was also significant, $F(2.87, 189.73) = 9.03$, $p < .001$, $\eta_p^2 = .120$. Bonferroni-corrected multiple comparisons showed that the proportions of switches in the first four periods were all higher than in the sixth period, ps $< .005$. There was no detectable difference between the fifth and sixth periods, $p = .072$. It appeared that interleaving gradually levelled off during self-regulated learning (see Figure S2). The interaction between group and period was not significant, $F(2.87, 189.73) = 1.02$, $p = .384$, $\eta_p^2 = .015$, suggesting that the difference in interleaving preference between the two groups remained steady across the study phase.

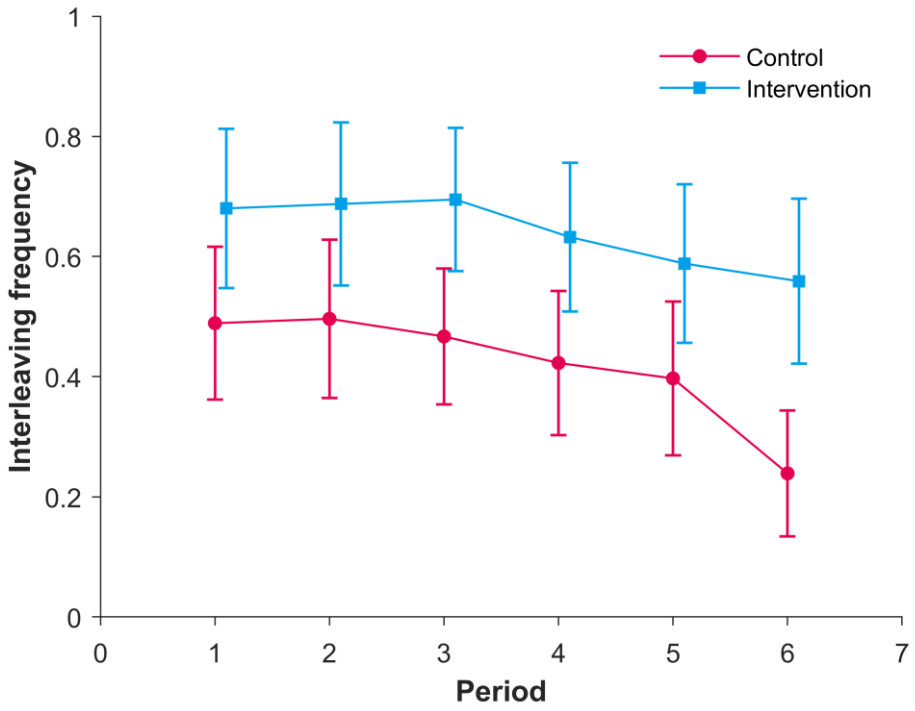


Figure S2. Interleaving frequency in the two groups as a function of time period in Experiment 2. Error bars represent 95% CI.

Interleaving frequency and distance as predictors of classification performance

The correlations between interleaving frequency, interleaving distance and classification performance in the post-intervention task are reported in Table S1. In the control group, interleaving distance positively predicted inductive performance. Interleaving frequency was also positively associated with classification accuracy, although the correlation was not statistically significant. With respect to the intervention group, both interleaving frequency and distance showed positive yet non-significant correlations with classification performance. After collapsing data across groups, the correlation between interleaving frequency and classification accuracy was marginally significant and the correlation between interleaving distance and classification accuracy reached significance.

Interleaving effect and magnitude of the instructional intervention effect

Participants in the intervention group were divided into two subgroups by their actual performance in the pre-intervention task. Only four out of 34 participants' classification accuracy did not demonstrate an interleaving effect. The numerical tendency in Table S2 suggests that it is more difficult for participants who did not show an interleaving effect to correct their metacognitive illusions and employ interleaving during self-regulated learning than those who did experience an interleaving effect.

Experiment 3

Exact wording of the instructions

The intervention instructions were adapted according to each participant's test performance in the pre-intervention task as follows:

For participants who correctly classified more paintings in the large interleaved than in the other two conditions, the instructions were as follows: *As shown in the left figure, you correctly classified X/24 (X%) paintings for artists studied via the blocked schedule, X/24 (X%) paintings for artists studied via the small interleaved schedule, and X/24 (X%) paintings for artists studied via the large interleaved schedule. Overall, your performance suggests that the interleaving strategy is superior to the blocking strategy. Indeed, prior psychological research has found that interleaving is normally more efficient than blocking for inductive learning. For instance, Kornell and Bjork (2008) observed that classification accuracy in the interleaved condition was about 27% higher than that in the blocked condition (see the middle figure).*

Moreover, your performance further indicates that large interleaving is superior to small interleaving. This is consistent with previous findings that mixing exemplars from more categories (large interleaving) is more effective than mixing exemplars from only a few categories (small interleaving). Indeed, Birnbaum et al. (2013) found that participants assigned to a large interleaving condition in which exemplars were intermixed across all 16 categories performed substantially better than those assigned to a small interleaving condition in which exemplars were intermixed across only 4 categories (see the right figure). Altogether, large interleaving is the most effective strategy for inductive learning.

For participants who did not show the best performance in the large interleaved condition, the instructions were as follows: *As shown in the left figure, you correctly classified X/24 (X%) paintings for artists studied via the blocked schedule, X/24 (X%) paintings for artists studied via the small interleaved*

schedule, and X/24 (X%) paintings for artists studied via the large interleaved schedule. Your performance in these 3 conditions is not consistent with previous research findings. Firstly, prior psychological research has found that interleaving is normally more efficient than blocking for inductive learning. For instance, Kornell and Bjork (2008) observed that classification accuracy in the interleaved condition was about 27% higher than that in the blocked condition (see the middle figure).

Moreover, it has been established that mixing exemplars from more categories (large interleaving) is more effective than mixing exemplars from only a few categories (small interleaving). Specifically, Birnbaum et al. (2013) found that participants assigned to a large interleaving condition in which exemplars were intermixed across all 16 categories performed substantially better than those assigned to a small interleaving condition in which exemplars were intermixed across only 4 categories (see the right figure). Although your performance is not entirely compatible with previous findings, it might just be a result of random variation. Altogether, prior studies demonstrated that large interleaving is the most effective strategy for inductive learning.

Exploratory analyses

Interleaving frequency as a function of time period

To explore learners' strategy regulation during the self-regulated learning phase, 80 discretionary choices in the post-intervention task were divided into 8 periods. Interleaving frequency was calculated as a function of time period and group. A two-way mixed-factor ANOVA was performed with Greenhouse–Geisser corrections. Group was entered as the between-subjects factor and period as the within-subjects factor. There was a main effect of group, $F(1, 64) = 14.47$, $p < .001$, $\eta_p^2 = .184$. The main effect of period was also significant, $F(2.94, 188.24) = 5.24$, $p = .002$, $\eta_p^2 = .076$. As shown in Figure S3, learners' tendency to interleave was highest at the beginning of the study phase and gradually decreased. Specifically, pairwise comparisons between each pair of periods showed that the proportion of switching in the first 6 periods were all significantly higher than that in the eighth period with Bonferroni-adjusted ps $< .038$. The interaction term was not significant, $F(2.94, 188.24) = 0.24$, $p = .862$, $\eta_p^2 = .004$, suggesting the intervention effects on strategy usage held constant over the course of the study phase.

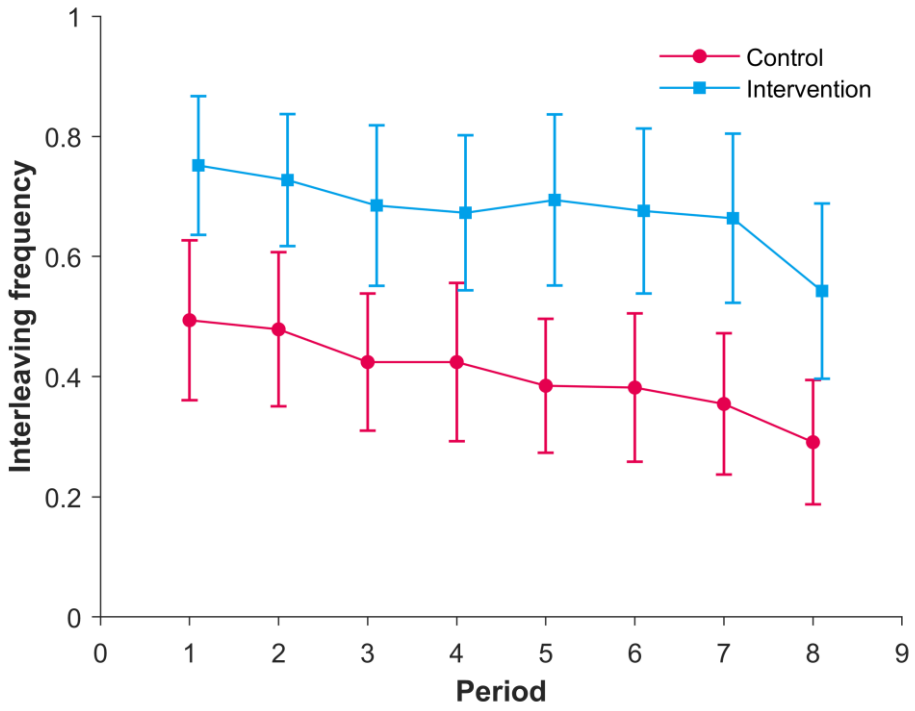


Figure S3. Interleaving frequency in the two groups as a function of time period in Experiment 3. Error bars represent 95% CI.

Interleaving frequency and distance as predictors of classification performance

With respect to the potential predictors of classification performance, the correlation between interleaving frequency and classification accuracy in the post-intervention task was computed, as well as the correlation between interleaving distance and classification accuracy (see Table S1). For the control group, only interleaving frequency positively correlated with classification performance and interleaving distance had a numerically positive correlation with classification performance. For the intervention group, both interleaving frequency and interleaving distance were positively associated with test performance. After collapsing data across groups, the correlation between interleaving frequency and classification accuracy was significant and the same was true for the correlation between interleaving distance and classification accuracy.

Mediation analysis

Experiment 3 revealed a significant intervention effect on the classification performance in the post-intervention task. To determine whether this effect stemmed from changes in participants' strategy employment after the intervention, a mediation analysis was conducted. Due to the high collinearity between interleaving frequency and interleaving distance (see Table S1), these metrics were combined into a single measure of strategy use by converting both to *z*-scores and subsequently calculating the average of these *z*-scores. The mediation effect was estimated using the bootstrap method, which involved generating 5,000 bootstrap samples to ensure a robust analysis. The 95% bootstrap confidence intervals and partially standardized coefficients are reported alongside the estimated coefficients.

As previously mentioned, the total effect of instructional intervention on classification performance in the post-intervention task was significant, $c = 0.07$ [0.002, 0.14], $c_{ps} = 0.50$. Critically, there was an indirect effect of intervention on classification performance through strategy usage, $ab = 0.04$ [0.01,

0.08], ab_{ps} = 0.29. Moreover, the direct effect was minimal, c' = 0.03 [-0.04, 0.10], c'_{ps} = 0.20, suggesting that the intervention effect on learning outcomes was completely mediated by enhanced self-employment of the interleaving strategy (see Figure S4).

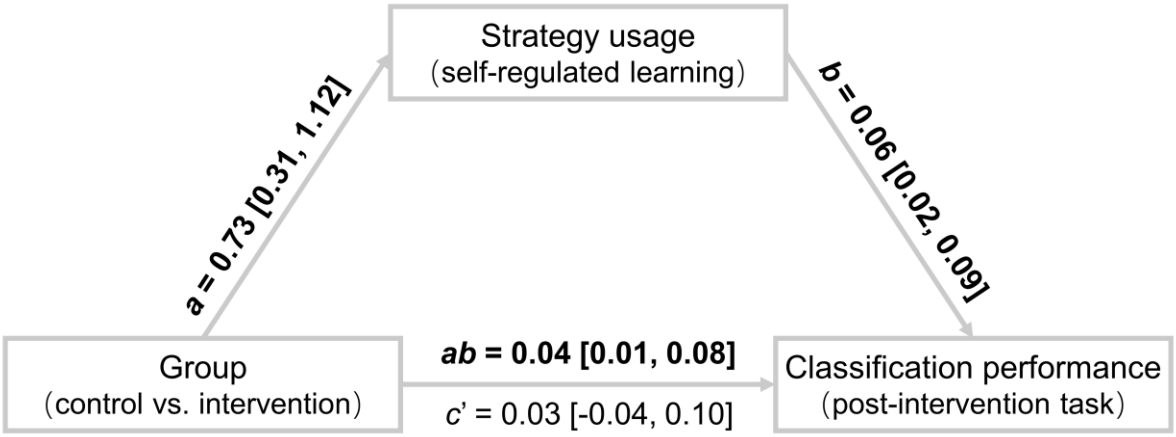


Figure S4. Results from the mediation analysis, with strategy usage—combined through interleaving frequency and interleaving distance—as the mediator. All CIs are 95% bootstrap CIs based on 5,000 samples.

References

Kensinger, E. A., & Corkin, S. (2003). Effect of negative emotional content on working memory and long-term memory. *Emotion*, *3*(4), 378-393. https://doi.org/10.1037/1528-3542.3.4.378

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

Lu, X., Penney, T. B., & Kang, S. H. K. (2021). Category similarity affects study choices in self-regulated learning. *Memory & Cognition*, *49*(1), 67-82. https://doi.org/10.3758/s13421-020-01074-w

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. https://doi.org/10.1037/xge0000177