Journal of Applied Research in Memory and Cognition 10 (2021) 589–602
Accepted by the outgoing editorial team

Contents lists available at ScienceDirect

Journal of Applied Research in Memory and Cognition

journal homepage: www.elsevier.com/locate/jarmac

Empirical Article

The Robustness of the Interleaving Benefit

Veronica X. Yan *
Educational Psychology, The University of Texas at Austin, USA

Faria Sana
Psychology, Athabasca University, Canada

Interleaving examples of to-be-learned categories, rather than blocking examples by category, can enhance learning. We examine the reliability of the interleaving effect between- (Experiments 1 and 2) and within-participants (Experiment 3). As a between-participant effect, we examined a broad spectrum of working memory by both measuring individual capacity and manipulating the task demand. Findings reveal a robust interleaving effect across the spectrum, eliminated only at the lowest and highest ends, but never reversed. In Experiment 3, we used an empirically defined source of potential heterogeneity by examining whether the size of the interleaving benefit a participant experiences on one set of stimuli predicts the size of the interleaving benefit that same participant experiences on two other sets of stimuli. It did not, with only a very small correlation between the two more similar stimuli sets. Taken together, these results add to the burgeoning literature on the robustness of the interleaving benefit.

Keywords: Interleaving, Category learning, Sequencing, Spacing, Working memory

General Audience Summary

The interleaving effect is the counterintuitive finding that studying or practicing multiple concepts in a mixed-up order leads to better learning than does focusing on one concept at a time. This interleaving benefit has been shown to apply to a wide range of learning tasks, from learning motor skills to more cognitive concepts such as recognizing the painting styles of different artists or solving mathematics problems. However, because most studies focus on average effects, less is known about how the interleaving effect varies between individuals. Individual differences are practically important—as a teacher, you would not want to apply interleaving broadly if it would help only some students and put others at a disadvantage. In fact, in the motor skills literature, there is some evidence that incorporating interleaved practice is not as effective for complex skills or for novice learners—that initial blocked practice is required before incorporating interleaved practice. We mimic low- and high-complexity by varying the task demands, and then examine how learners with low and high working memory capacity benefit from interleaving. We did not find a reversal where blocking was more effective. Finally, we ask if there is a subset of learners for whom interleaving is reliably not beneficial across multiple sets of learning materials. There was not. These findings together provide deeper insight into the generalizability and robustness of the interleaving effect. It adds to the literature, showing that interleaving does not just promote learning across different materials, but also across different learners.

Open up almost any textbook or examine almost any lesson plan, and you will notice that concepts are most often taught in discrete blocks. In commonly adopted seventh-grade mathematics textbooks, for example, Rohrer, Dedrick and Hartwig

* Correspondence concerning this article should be addressed to Veronica X. Yan, The University of Texas at Austin, Austin, TX, USA. Contact: veronicayan@austin.utexas.edu (V.X.Y.).

(2020) found that 91% of practice problems were structured in such a blocked manner. This observation sounds obvious and trite—it makes sense that concepts are taught and practiced one at a time before we move onto the next. The alternative option—to mix up, or "interleave" the study of different concepts—would surely feel disorganized and confusing to learners. And yet, a surprising number of empirical studies have shown that interleaving practice of different concepts can lead to better learning than studying one concept at a time (see Brunmair & Richter, 2019 for a recent meta-analysis).

Generalizability of the Interleaving Effect

The interleaving benefit has been demonstrated across many different types of concepts and age groups. For example, interleaved study or practice has been demonstrated to benefit the learning of perceptual categories, such as artists' painting styles (Kang & Pashler, 2012; Kornell & Bjork, 2008), butterfly species (Birnbaum et al., 2013), and chest radiographic patterns (Rozenshtein et al., 2016). Interleaved study has also been shown to benefit the learning of cognitive concepts, such as mathematics formulae (Rohrer, 2012; Taylor & Rohrer, 2010), the application of non-parametric statistics to different research designs (Sana et al., 2017), the diagnosing of clinical disorders in case patients (Zulkiply et al., 2012), and the classification of organic chemistry compounds (Eglington & Kang, 2017). Finally, there is also a large body of motor skills research demonstrating the benefit of interleaving with skills ranging from simple sequences (Simon & Bjork, 2001) to large, complex movements in sports (e.g., badminton, Goode & Magill, 1986; volleyball, Jones & French, 2007; baseball, Hall et al., 1994; golf; Brady, 2016), to fine movements such as playing musical instruments (Abushanab & Bishara, 2013) and tying knots (Ollis et al., 2005).

The benefit of interleaving has also been shown across different age groups. While most of the research has been conducted with college students, and young- and middle-aged adults, interleaving has also been shown to benefit learning for children (Nemeth et al., 2019; Taylor & Rohrer, 2010; Vlach et al., 2008), and for older adults (Kornell et al., 2010; Lin et al., 2012, 2016).

Boundary Conditions: When is Interleaving Not Better than Blocking?

Characteristics of learning tasks. One of the key theories as to why interleaving may benefit category learning is that an interleaved schedule juxtaposes examples from different categories and draws learners' attention to the features that discriminate between categories while blocking by categories draws learners' attention to the features that are shared within a category (sequential-attention theory, Carvalho & Goldstone, 2017; discriminative-contrast theory, Kang & Pashler, 2012). Following this line of logic, multiple studies have shown that inserting other tasks between examples to disrupt learners' ability to easily identify discriminating features between categories can attenuate or even eliminate the interleaving benefit (Birnbaum et al., 2013; Sana et al., 2017).

Under this sequential-attention theory, the implication is that if categories are already highly discriminable, it may be more important to direct learners' attention to the within-category similarities. If so, then blocking benefits should be obtained. Indeed, Carvalho and Goldstone (2014) show that when categories are highly similar (i.e., high need for discrimination), interleaving is more beneficial, but when categories are dissimilar, blocking is more beneficial (cf. Foster et al., 2019).

Characteristics of individual learners. Another key theory as to why interleaving may benefit category learning is that interleaving inherently involves spacing (i.e., distributed learning). In the motor skills literature, researchers have found that interleaving does not always benefit learning when difficulty level is high, either because the skill is complex or because the learners are novices (Magill & Hall, 1990; Porter & Magill, 2010; Shea et al., 1990; Wulf & Shea, 2002). For example, in studies comparing blocked and interleaved practice of tennis and golf strokes, the interleaving effect was larger for experts than it was for novices (Guadagnoli et al., 1999; Hebert et al., 1996).

Unlike in motor literature, where prior skill set and task difficulty are potential moderators of the interleaving benefit, in cognitive literature, "individual difference" has mostly been operationalized as working memory capacity[1]. Four published studies have examined the interaction between working memory capacity and the interleaving effect. It may be hypothesized that interleaved schedules increase working memory demands due to the cognitive load required to hold the features of several different categories (versus focusing on just the features of one category) in working memory at the same time. Hence the interleaving effect may be attenuated or even eliminated for individuals with lower working memory capacity. However, studies have not found this effect (Guzman-Munoz, 2017; Sana et al. 2017; Sana et al., 2018; Wang et al., 2020).

While Guzman-Munoz (2017) and Sana et al. (2017, 2018) examined only individual's working memory capacities, Wang et al. (2020) also manipulated working memory load, by comparing the interleaving benefit under single and dual-task conditions. In studies that examined working memory capacity, the category-learning task load may not have been sufficient to adversely affect learners with low working memory capacity. Moreover, the samples from these studies were drawn from relatively homogenous and high-performing populations (college-students), which alone may be insufficient to examine the interleaving effect across the broader range. In the study that also manipulated working memory load of the task (Wang et al., 2020), the dual-task manipulation involved a numerical Stroop task in which participants were shown two

[1] Two studies (Kirk-Johnson, Galla, & Fraundorf, 2019; Yan, 2014) have also examined individual difference in terms of intelligence mindsets (cf. Dweck, 2006). Neither found that a learner's belief about the malleability of intelligence is related to the interleaving effect.

numbers of differing physical size and numeric value before each painting, and then asked to report the side on which the number had the larger size of value after each painting. This dual-task manipulation lowered overall performance but did not interact with schedule. This specific manipulation, however, also served to disrupt the juxtaposition between consecutive paintings, possibly changing the nature of the task (see Birnbaum et al., 2013; Kang & Pashler, 2012; Sana et al., 2017). In the present studies, we address these issues.

The Present Studies

A majority of the studies so far have typically reported the average effects of interleaving. In the present study, we ask whether there is heterogeneity in sequencing effects, even with stimuli that typically show interleaving benefits on average. Given a set of material that a teacher knows usually benefits from interleaving, should the teacher be worried that it interleaving might lead to a rich-gets-richer effect and leave struggling students behind? In Experiments 1 and 2, we measured working memory capacity as an individual difference, and we experimentally manipulated the working memory load of the category-learning task. Hence, even if participants themselves are relatively homogeneous, there would be a wide range of cognitive task load. This design allows us to explore the interleaving effect along a broader spectrum of working memory than only using working memory capacity as a covariate or only manipulating task cognitive load.

In Experiment 3, we took a different approach to examine the heterogeneity of the interleaving effect. When a study shows an overall interleaving benefit, there may be a subset of participants for whom the blocked learned categories are better classified on the final test. Within-participant stability of the interleaving benefit is important because if there are individual differences in who learns better through blocking, then organizing everything in an interleaved manner would put some learners at a particular disadvantage. One possible way of identifying individual differences is to test covariates such as working memory load (as in Experiments 1 and 2), learners' prior knowledge (Hebert et al., 1996), learners' categorizing strategies (e.g., McDaniel et al., 2014), learners' beliefs (Kirk-Johnson et al., 2019; Yan, 2014), and so on. However, this approach is very tedious, requiring that experimenters somehow hit on just the right covariate and calibrate the design well enough to sample an appropriate range. Another possible way of looking for individual differences is to present participants with varying stimuli sets of blocked and interleaved study schedules to examine whether there is a stable subset of participants for whom interleaving is not beneficial. Hence, in Experiment 3, we presented participants with three study-test cycles to examine how consistent, within-participants, the efficacy is of the blocked and interleaved study schedules. This design allowed us to test whether the advantage (or lack thereof) of interleaving when learning one set of stimuli is related to the advantage (or lack thereof) of interleaving when learning the other two sets of stimuli. The open data for all three of the experiments can be found at http://osf.io/9tn2j/.

Experiment 1

In Experiment 1, we explored working memory (WM) as a moderator of the interleaving benefit in two ways: as an individual difference using a WM span task as our indicator of between-participant differences, and as an experimentally manipulated variable using a dual-task paradigm. In real life, these are important ways of looking at the effect—examining individual WM capacity can tell us about whether interleaving works across participants; examining conditions with and without a WM load can tell us whether interleaving is likely to work across different contexts in which learners are under more or less load (e.g., test anxiety, distracting environments). But more importantly, such a design would yield a dataset in which we could better capture a broader range of working memory load (with near floor, for those with lower working memory capacity in the high load condition, and near ceiling effects, for those with higher working memory capacity in the low load condition). In turn, this range would allow us to examine the robustness of the interleaving benefit across the WM load spectrum.

We not only expected to replicate prior findings on the interleaving effect, but also sought to manipulate WM load in a way that does not disrupt a key mechanism of the interleaving effect, namely discriminative contrast. In our design, participants maintained a series of digits in their WM for a longer period of time—the duration of six sequentially presented paintings versus the duration of one painting presentation as in Wang et al. (2020). This did not only increase the WM load, but also minimized the disruption of compare and contrast processes between consecutive paintings.

Method

Participants and design. A total of 143 undergraduate students (103 females; Mean age = 18.66, *SD* = 1.62, range 18–31) participated in exchange for course credit. Presentation schedule (blocked vs. interleaved) was manipulated within-subjects and the working memory load (low load vs. high load) was manipulated between-subjects. Participants were randomly assigned to either the low WM load condition ($n = 68$) or to the high WM load condition ($n = 75$).

Materials. The stimuli, taken from Kornell and Bjork (2008), were landscapes or skyscapes by 12 artists: Georges Braque, Henri-Edmond Cross, Judy Hawkins, Philip Juras, Ryan Lewis, YieMei, Marilyn Mylrea, Bruno Pessani, Ron Schlorff, Georges Seurat, Ciprian Stratulat, and George Wexler. See the first row of Figure 1 for examples.

Because WM capacity measures differ widely across studies, we chose a complex span task to assess WM in the present study in line with other studies assessing the relation between WM and higher-order cognition (e.g., Conway et al., 2005). Complex span tasks represent WM as a multi-faceted system that captures variance from different processes subsumed under WM, such as attentional control and controlled search from long-term memory (Unsworth & Engle, 2007). In this study, therefore, participants performed an automated version of the operation span task (OSPAN; Kane et al., 2004; Unsworth

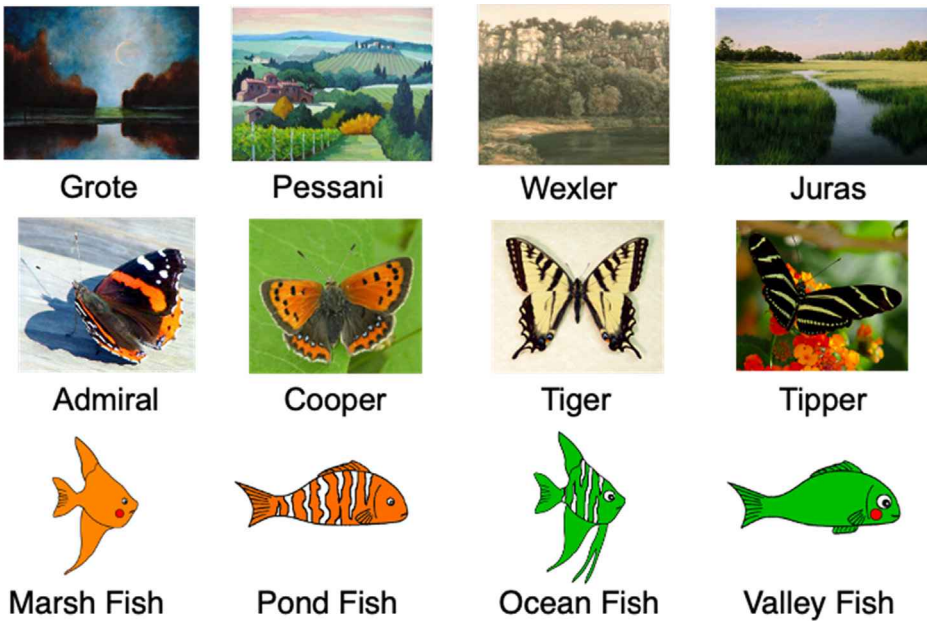

Figure 1. Examples of stimuli across the three experiments. The first row illustrates artists' painting styles (Experiments 1 and 3), the second row illustrates butterfly species (Experiment 3), and the third row illustrates cartoon fish (Experiments 2 and 3).

et al., 2005). Briefly, the OSPAN required participants to solve a series of math problems while trying to remember a sequence of unrelated letters, ranging from three to seven letters in length[2]. At the completion of the task, five scores were calculated. The OSPAN score—the measure used herein and the most common one used to index WM capacity (see Conway et al., 2005)—is the sum of all letters from the letter sets that were recalled completely in the correct order. Full details of task structure and timing can be found in Unsworth et al. (2005).

Procedure. Participants were told that they would be working on two tasks: a digit memory task and a paintings task. For the digit memory task, they were told that they would be shown three or four two-digit numbers to memorize, and then be asked to recall them. Each sequence would only be tested only once, and hence could be forgotten as soon as that recall trial happened. For the paintings task, they would be shown paintings by 12 different artists and their task was to learn to recognize their painting styles, such that they would be able to identify (from a list of names) the artist responsible for new, never-before-studied paintings on the final test.

Demonstration. To give participants a concrete understanding of what these two tasks involved, we first gave them a brief demonstration. In the low WM load (i.e., control) condition, participants were presented with three two-digit numbers (e.g., 84, 52, 19) for four seconds. They were then given 10 seconds to recall the three numbers. Next, they studied six paintings for three seconds each. They studied two paintings by each of three well-known artists: Salvador Dali, Vincent Van Gogh, and Pablo Picasso. The two paintings by a given artist were always presented consecutively in the demonstration phase. The artist's surname was presented underneath each painting. In the high WM load condition, the demonstration procedure was similar, but with one difference: Instead of being asked to recall the three numbers immediately after study, they were only asked to recall the numbers after studying the six paintings.

Name pre-training. To eliminate any differential effects that name learning might have on participants, they were first familiarized with the artists' names before they began studying the paintings. In the familiarization phase, participants were given 45 s to study the 12 names, and then they engaged in three cycles of two-letter-stem cued recall tests (with immediate feedback). By the third cycle, all participants could recall the 12 artists' names with high accuracy (allowing for spelling errors).

Study phase. During the study phase, six paintings by each artist (for a total of 72) were presented sequentially for 3 s each, with the artist's name presented below each painting. For each individual, the twelve artists were randomly divided into the blocked and interleaved conditions. The paintings were presented in the following sequence: B I I B B I I B B I I B, where each "B" refers to a blocked set of six paintings and where each "I" refers to an interleaved set of six paintings. Each blocked set consisted of six paintings by the same artist. Each interleaved set consisted of one painting by each of the six interleaved artists.

In between each set of six paintings, participants were asked to complete a digit memory and recall task. Participants were shown either three or four two-digit numbers (e.g., 15, 85, 19, 80) and allowed to study the digit sequence for four seconds. Participants were randomly assigned into the low WM

[2] We chose the OSPAN because unlike other complex span tasks, such as the Reading span task, the OSPAN task was specifically designed to tap into domain-general abilities instead of domain-specific abilities, and therefore is less sensitive to influences of reading or language ability (e.g., Turner & Engle, 1989).

load or the high WM load conditions. In the low WM load condition, they were asked immediately to recall the numbers before moving onto studying the next set of six paintings. Participants were allotted 10 seconds to recall the digits before the program moved on. In the high WM load condition, however, they were not asked to recall the numbers until the next set of six paintings had been studied. Hence, they had to hold them in mind during the study of the paintings. Once they recalled the previous sequence of numbers, they were immediately given their next sequence of numbers to memorize before studying the next set of six paintings. For both control (i.e., low WM load condition) and digit rehearsal condition, the first set of numbers was always given to participants before the first set of six paintings; hence, everyone studied and recalled a total of 12 sequences of numbers.

Test phase. After completion of the study phase and a 90-s distraction period of playing Tetris, the final test was administered. Participants were shown four new paintings by each artist and asked to select, from a list of names, the artist responsible for each painting (for a total of 48 test items). The test images were presented sequentially in four sets, with each set containing one new painting by each of the 12 artists, randomly ordered. As soon as participants made their choice by clicking on an artist's name, the next painting was presented. The test was self-paced and included no feedback.

Post-test. Following the test, the two different schedules were described to participants, and they were asked which schedule they thought would be more effective for the learning of the artists' painting styles (interleaved, blocked, or equally effective). Finally, participants performed the OSPAN task.

Results and Discussion

Manipulation checks. We first checked whether our WM load was meaningfully experienced by participants. First, we examined the accuracy on the interspersed digit memory task. Participants in the high load condition ($M = 46.57$, $SD = 0.25$) performed significantly worse than those in the control condition ($M = 64.82$, $SD = 17.34$), $t(131.22) = 5.05$, $p < .001$, $d = 0.84$. At the end of the classification task, participants in the high load condition were less likely to report than they tried to verbalize distinguishing features between the artists ($M = 2.28$, $SD = 1.00$) than those in the low load condition ($M = 2.93$, $SD = 1.00$), $t(139.06) = 3.83$, $p < .001$, $d = 0.64$. Participants in the high load condition also rated that it was more difficult to come up with rules that helped them distinguish between artists ($M = 2.68$, $SD = 0.94$) than participants in the low load condition ($M = 3.00$, SD = 0.83), $t(138.57) = 2.11$, $p = .036$, $d = 0.35$.

As the OSPAN task was administered at the end of the study, we also checked the possibility that participants in the high load condition might have performed more poorly if they were more fatigued than the participants in the low load condition. However, an independent t-test revealed that there was not a significant difference in the OSPAN scores of those in the low load condition ($M = 49.24$, $SD = 17.00$) and those in the high load condition ($M = 48.37$, $SD = 20.81$), $t(139.53) = 0.27$, $p = .79$, $d = 0.05$. The mean OSPAN score was 48.78 ($SD = 19.03$, median = 50), with a range of 3–96.

Classification test performance. The average final classification test scores by schedule and WM load condition are presented in Table 1. The top panel of Figure 2 shows the distribution of individuals' interleaved and blocked score differences (averaged across stimulus types) in Experiment 1.

We conducted a linear mixed effects model, predicting accurate classification on the final test from WM load condition, schedule condition, participants' individual OPSAN score (standardized) and the interaction between the three predictors. The artist category and the individual ID were also entered as random effects. The summary statistics of this regression analysis are presented in Table 2.

Results revealed three main effects: Participants with higher WM spans scored higher than those with lower WM spans; participants in the low WM load (control) condition scored higher than those in the high WM load condition; and the artists whose paintings had been studied interleaved were more likely to be correctly classified than those whose painting had been studied blocked. In other words, we replicate patterns from prior studies showing that WM matters, and that there is an interleaving benefit.

There were also significant interactions. There was a significant OSPAN × schedule interaction—interleaving was particularly likely to benefit the learning of those with low WM capacity. This two-way interaction, however, is qualified by a significant three-way interaction between individual WM capacity, WM load of the task, and schedule. The three-way interaction is illustrated in Figure 3. One way to interpret this three-way interaction is that under conditions of low WM load (right panel), interleaving may be particularly beneficial for those with lower WM capacity (it eliminated the difference between low and high WM capacity); however, under high WM load (left panel), interleaving may be particularly beneficial for those with higher WM capacity.

However, we believe that Figure 3 indicates a subtly different interpretation. Note that interleaving benefits are generally obtained unless performance is very low (low WM capacity, high load) or very high (high WM capacity, low load). Moreover, under the low load condition, there is no difference

Table 1
Classification Test Performance by Working Memory Load and Schedule Condition in Experiments 1 and 2

Experiment	WM load condition	Blocked mean (SD)	Interleaved mean (SD)	Cohen's d (95% CI)
1	High load	0.25 (0.17)	0.41 (0.22)	0.78 (0.44, 1.11)
	Low load	0.42 (0.21)	0.58 (0.23)	0.72 (0.37, 1.06)
2	High load	0.38 (0.28)	0.52 (0.33)	0.45 (0.13, 0.77)
	Low load	0.43 (0.31)	0.61 (0.31)	0.58 (0.24, 0.91)

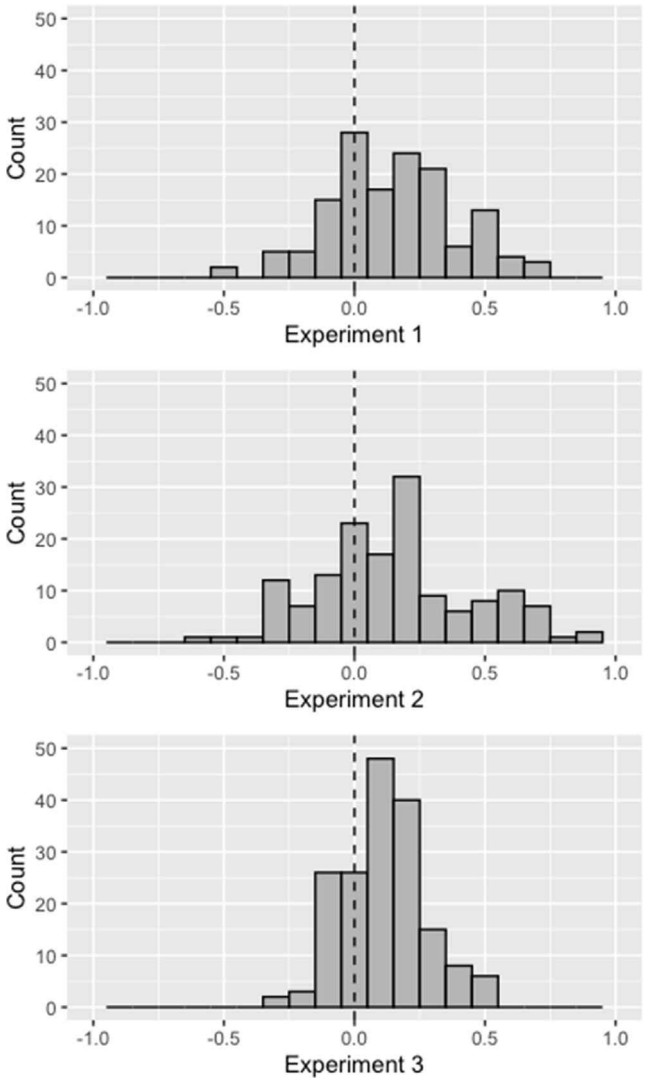

Figure 2. Histogram of interleaving benefits in Experiments 1–3. The x-axis represents the difference between the averaged interleaved score and the averaged blocked score. Values greater than zero on the x-axis (indicated by the dashed line) represent an overall interleaving benefit.

between those with low and high WM capacity. We hence believe that the lowest accuracies reflect a floor effect, while the highest accuracies reflect a functional ceiling effect, given this particular task context (of course, performance could likely get closer to 100% if, for example, participants were given more examples, more time, more repetitions, and so on). In other words, we believe that our data capture the broad range of possible performance in this task and rather than overinterpreting the interactions, the data seem to indicate an interleaving benefit across the majority of this range.

Metacognitive judgment. Following the final classification test, the different schedules were described to participants, and they were asked which schedule they thought would be more effective for the learning of the artists' painting styles (interleaved, blocked, or equally effective). The responses are shown in Table 3. Participants overwhelmingly thought that blocking was more effective for their own learning.

Experiment 2

In Experiment 1, the stimuli were artists' paintings. Across the spectrum of WM load (except for when the data were subject to floor and ceiling effects), we found interleaving benefits. In Experiment 2, we replicate the design of Experiment 1, but used different categories—artificially created cartoon fish. We chose these rule-based fish because with stimuli that have clearly defined rules, a ceiling effect in the range of 60–70% (as we think we may have with the artist stimuli) is much less likely.

Method

Participants and design. One-hundred and fifty undergraduates (mean age = 19.32, *SD* = 1.74, range 18–30 years; 122 female, 27 male, 1 declined to disclose) were recruited and compensated with partial course credit. Presentation schedule (blocked vs. interleaved) was manipulated within-subjects and the WM load (low load vs. high load) was manipulated between-subjects. Participants were randomly assigned to either the low load condition (n = 72) or to the high load condition (n = 78).

Materials. The fish stimuli were a subset of those used in Yan et al. (2014, poster). See the third row of Figure 1 for examples of these materials. Participants studied six categories. The total stimuli set consisted of 16 examples per fish category. For each individual, a randomly selected set of 12 examples were used in the study phase and the remaining four examples were used in the test phase. Each fish image was comprised of seven binary features: body shape (long vs. tall), color (orange vs. green), pattern (solid vs. striped), eye shape (small vs. large), mouth (smile vs. frown), bottom fin (present vs. absent), and red cheek dot (present vs. absent). Each category was

Table 2
Regression Summary Statistics, Predicting Classification Test Accuracy in Experiment 1

Predictor	*b*	*SE*	*df*	*t*	*p*
Intercept	0.42	0.04	24.49	11.76	<0.001
OSPAN (z-scored)	0.06	0.02	184.87	2.55	0.01
Schedule	0.16	0.02	6710.16	10.39	<0.001
WM-load	−0.16	0.03	184.80	−5.52	<0.001
OSPAN × Schedule	−0.06	0.02	6709.70	−3.50	<0.001
OSPAN × WM-load	−0.05	0.03	184.90	−1.56	0.12
WM-load × Schedule	−0.01	0.02	6709.21	−0.63	0.53
OSPAN × Schedule × WM-load	0.12	0.02	6710.00	5.33	<0.001

Note. The reference level for schedule is blocked; the reference level for WM-load is low-load. Marginal R^2 = 0.064; 6864 observations, 143 participants, 12 artists.

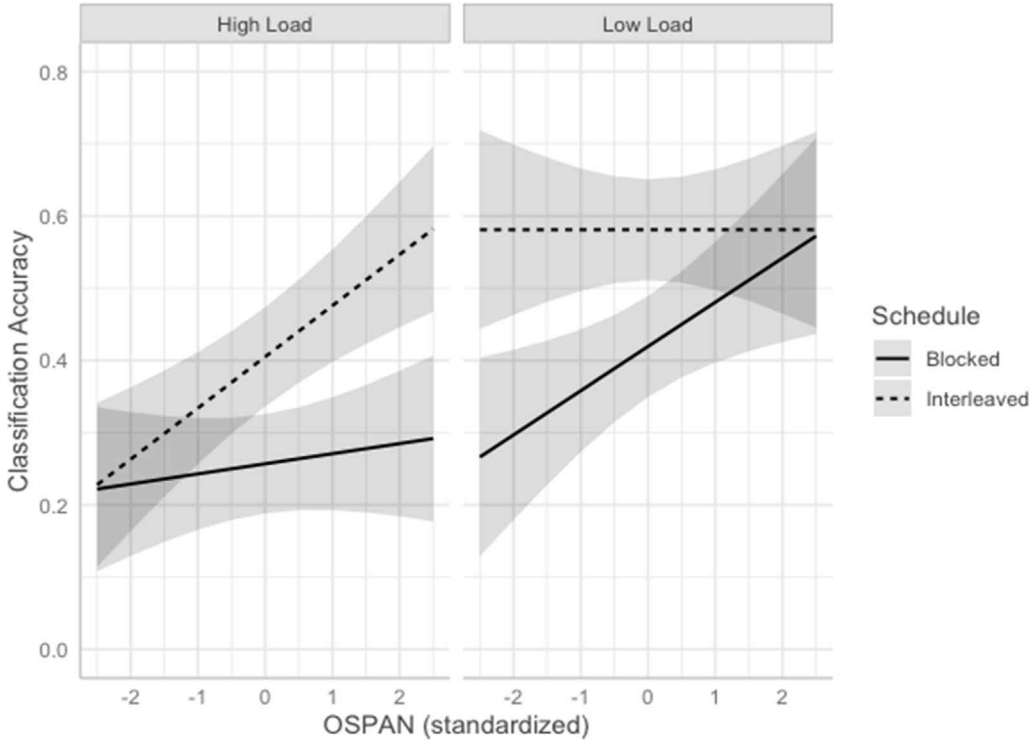

Figure 3. Results of regression analysis for Experiment 1, predicting classification test performance from schedule, WM load condition, and WM capacity. Individual and category were entered as random effects. The shading around the lines reflect 95% confidence intervals.

defined by a particular combination of body shape, color, and pattern; the other features varied randomly. The same OSPAN task as that used in Experiment 1 was used to measure individuals' WM capacity.

Procedure. The procedure was similar to that of Experiment 1. Participants were told that they would be working on two tasks: a digit memory task and a task involving recognizing different cartoon fish "species." The digit memory task was the same as that of Experiment 1.

Demonstration and name pre-training. The demonstration and name pre-training procedure was very similar to that of Experiment 1. There were only two differences. First, in the demonstration, instead of seeing paintings by famous artists, they were simply shown a message that said, "[Image will appear here]" in the center of the screen with one of six of the fish category names (e.g., Bay Fish) below the message. They saw each fish category name once, in a random order. Second, in the name pretraining stage, when participants were first allowed to study all six fish category names, they were given 30 s (instead of 45 s).

Study phase. Participants studied 12 examples of each of the six fish categories; examples were presented sequentially for 3 s each, with the fish category name presented below each example. For each individual, the six fish were randomly assigned into the blocked and interleaved conditions; the 72 stimuli were assigned to six sets: three blocked sets (12 examples of one fish category per set) and three interleaved sets (four examples of each of three fish categories per set). The order of the six sets was randomized for each individual. Every six examples, participants were asked to complete a digit memory and recall task; this was the same as in Experiment 1.

Test phase. After completion of the study phase and a 90-s distraction period of playing Tetris, the final test was administered. Participants were shown four new examples of each fish category and asked to select, from a list of names, the category to which each example belonged (for a total of 24 test items). The test images were randomly ordered. As soon as participants made their choice by clicking on a fish category name, the next example was presented. The test was self-paced and included no feedback.

Post-test. Following the test, the two different schedules were described to participants, and they were asked which schedule they thought would be more effective for the learning of the fish species (interleaved, blocked, or equally effective). Finally, participants completed the OSPAN task.

Table 3
Participants' Metacognitive Judgments of Most Effective Schedule in Experiments 1 and 2

Metacognitive judgment	Experiment 1	Experiment 2
Interleaved	16 (11%)	26 (17%)
Blocked	124 (87%)	106 (71%)
Equally effective	2 (1%)	18 (12%)

Results and Discussion

Manipulation checks. We first checked whether our working memory load was meaningfully experienced by participants. First, we examined the accuracy on the interspersed digit memory task. Participants in the high load condition

(M = 28.50, SD = 15.92) performed significantly worse than those in the control condition (M = 84.76, SD = 12.24), t (143.42) = 24.37, $p < .001$, d = 3.96. At the end of the classification task, participants in the high load condition were less likely to report that they tried to verbalize distinguishing features between the fish categories (M = 3.50, SD = 1.91) than those in the low load condition (M = 4.19, SD = 1.98), t (146.12) = 2.18, p = .031, d = 0.36; but they did not rate it as more difficult to come up with the rules that helped them distinguish between the categories (M = 3.74, SD = 1.38) than participants in the low load condition (M = 3.94, SD = 1.57), t (141.78) = 0.83, p = .41, d = 0.14. There was also no difference in the number of correct rules (out of a total of seven, one for each feature) identified between the participants in the low load condition (M = 5.79, SD = 1.27) and participants in the high load condition (M = 5.55, SD = 1.22), t(146.07) = 1.18, p = .24, d = 0.19.

As the OSPAN task was administered at the end of the study, we also checked the possibility that participants in the high load condition might have performed more poorly if they were more fatigued than the participants in the low load condition. However, an independent t-test revealed that there was no significant difference in the OSPAN scores of those in the low load condition (M = 48.06, SD = 16.40) and those in the high load condition (M = 47.73, SD = 18.28), t(147.89) = 0.11, p = .91, d = 0.02. The mean OSPAN score was 47.89 (SD = 17.35, median = 49), with a range of 0–75.

Classification test performance. The average final classification test scores by schedule and WM load condition are presented in Table 1. The middle panel of Figure 2 shows the distribution of individuals' interleaved and blocked score differences (averaged across stimulus types) in Experiment 2.

We conducted a linear mixed effects model, predicting accurate classification on the final test from WM load condition, schedule condition, participants' individual OSPAN score (standardized) and the interaction between the three predictors. The fish category and the individual ID were also entered as random effects. The summary statistics of this regression analysis are presented in Table 4. Results revealed only one significant predictor of classification accuracy: those categories that were studied interleaved were significantly more likely to be correctly classified than those categories that were studied blocked. Individual WM capacity was a marginally significant predictor, and WM load was not a significant predictor.

There were also no significant interactions; nevertheless, to aid comparison with the results of Experiment 1, we present the three-way interaction (or lack thereof) visually in Figure 4. This figure shows that, similar to the results of Experiment 1, when test accuracy is at its lowest (low WM capacity, high WM load), there appears to be no benefit of interleaving, but this likely reflects a floor effect. Unlike the artists' paintings stimuli, however, test accuracy at the high end (high WM capacity, low WM load) does not appear to have hit a functional ceiling; this difference is likely the reason why we did not find the same statistical interactions as we did in Experiment 1. Despite the surface differences in significant (or not) interactions, we believe the take-away is the same: the interleaving effect is robust across the majority of the range.

Metacognitive judgment. Following the final classification test, the different schedules were described to participants, and they were asked which schedule they thought would be more effective for the learning of the artists' painting styles (interleaved, blocked, or equally effective). The responses are shown in Table 3. Participants overwhelmingly thought that blocking was more effective for their own learning.

Experiment 3

In Experiments 1 and 2, we measured individual differences in WM capacity and experimentally manipulated WM load during the category-learning task. Our results indicated a robust interleaving benefit across a large range of task difficulty (spanning from performance at floor to ceiling or near-ceiling). In Experiment 3, rather than identifying a specific type of individual difference, we used an empirically defined source of heterogeneity: the size of the interleaving benefit on an initial study-test cycle. In prior studies that have manipulated blocked and interleaved practice within-subjects, there are often a minority of participants whose final classification test performance scores do not reveal an interleaving benefit. Do these participants represent random noise or a special subset of the population for whom blocking might in fact be more effective than interleaving? That is, what is the test–retest reliability of this blocking benefit or this interleaving benefit? To examine this question, in Experiment 3, we presented participants with three study-test cycles and examined the consistency, within-participants, of the efficacy of the blocked and interleaved schedules.

Table 4
Regression Summary Statistics, Predicting Classification Test Accuracy in Experiment 2

Predictor	*b*	*SE*	*df*	*t*	*p*
Intercept	0.407	0.03	40.06	13.70	<0.001
OSPAN (z-scored)	0.049	0.03	264.2	1.94	0.054
Schedule	0.160	0.02	3458	7.84	<0.001
WM-load	−0.002	0.02	3472	−0.08	0.934
OSPAN × Schedule	0.011	0.02	3459	0.53	0.597
OSPAN × WM-load	0.011	0.02	3473	0.56	0.577
WM-load × Schedule	−0.005	0.03	3466	−0.16	0.870
OSPAN × Schedule × WM-load	0.004	0.03	3471	0.12	0.901

Note. The reference level for schedule is blocked; the reference level for WM-load is low-load. Marginal R^2 = 0.040; 3600 observations, 150 participants, 6 fish categories.

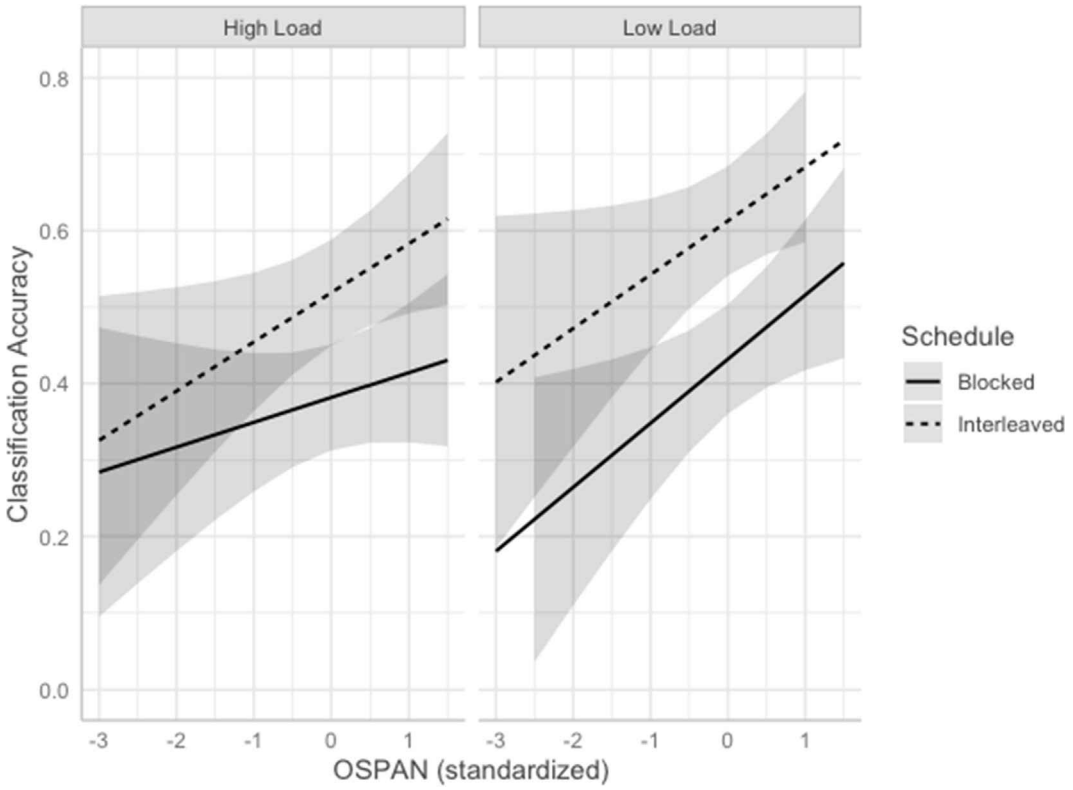

Figure 4. Results of regression analysis for Experiment 2, predicting classification test performance from schedule, WM load condition, and WM capacity. Individual and category were entered as random effects. The shading around the lines reflect 95% confidence intervals.

Method

Participants and design. One hundred and seventy-four participants (Mean age = 20.61, *SD* = 2.15, range 18–34) were recruited from the undergraduate psychology participant pool in exchange for course credit. Participants engaged in three study-test cycles, with the order of the category types (artists, butterflies, and fish) counterbalanced. Participants were randomly assigned to one of the counterbalancing conditions. Presentation schedule (blocked vs. interleaved) was manipulated within participants. For each individual, specific categories were randomly assigned to either blocked or interleaved condition, with the exception of the fish. Half of the fish categories had long trout-like bodies; the other half had triangular angelfish-like bodies. Instead of randomly assigning fish to blocked and interleaved categories, fish of the same body were always assigned to the same scheduling condition. The body-schedule pairing was counterbalanced across individuals.

Materials. We created three sets of to-be-learned material: artists' paintings, butterflies, and cartoon fish. Each set consisted of eight examples for each of eight categories (i.e., total of 64 examples). The artists materials consisted of landscape paintings by the following eight artists: Toni Grote, Judy Hawkins, Philip Juras, Lori McNamara, Marilyn Mylrea, Bruno Pessani, Ciprian Stratulat, and George Wexler. The butterflies included those used by Birnbaum et al. (2013) (additional examples had to be created to ensure there were enough for our design). The fish stimuli were taken from those used by Yan et al. (2014, poster). See Figure 1 for examples of these materials.

Procedure. Participants were informed that they would be completing three cycles of learning and testing, and that in each cycle, they would be learning something different. They were informed of what they were to learn in each cycle, and that the tests within each cycle would involve being shown new images and picking the appropriate category label from a list of names.

Demonstration and name pre-training. Before each study cycle began, they were introduced to what they were supposed to learn in that cycle. For example, if they were beginning a cycle of learning artists' paintings, they were told, "*You will be shown paintings by 8 different artists (8 paintings by each a total of 64 paintings). Your goal is to learn their painting styles. On a later test, you will be shown NEW paintings by the 8 artists and will have to identify, from a list of names, which artist painted that painting.*"

Study phase. Within each study phase, participants were shown one example at a time for 5 seconds each, with the name presented underneath the image. Four of the categories within each cycle were randomly assigned to the blocked condition; the remaining four were assigned to the interleaved condition. These examples were presented in a B I B I B I B I order or an I B I B I B I B order, where each "B" represents eight examples of the same category, and each "I" represents two different examples from each of four categories (randomized).

Table 5
Classification Test Performance by Schedule in Experiment 3

	Blocked score (SD)	Interleaved score (SD)	Cohen's *d* (95% CI)
Artists	0.49 (0.28)	0.60 (0.28)	0.45 (0.24, 0.66)
Butterflies	0.57 (0.31)	0.66 (0.33)	0.37 (0.17, 59)
Fish	0.49 (0.32)	0.63 (0.34)	0.53 (0.30, 0.73)
Overall	0.52 (0.30)	0.63 (0.32)	0.45 (0.24, 0.66)

Test phase. The test phase of a given study-test cycle was separated from the study phase by a 2-minute game of Tetris. Three new examples from each category were shown in the test phase, one example at a time. The order of the test examples were block randomized—that is, there were three test cycles, in which each cycle consisted of one example from each of the eight categories in a random order. The example was presented in the center of the screen with the category names presented as buttons underneath, arranged in two rows of four. The test phase was self-paced and participants clicked on a name to indicate their answer and move onto the next test image. They were not given feedback. After the final test image, the participants were informed that they had reached the end of the cycle, and would be starting on the next cycle. This procedure was repeated until they had completed all three study-test cycles.

Post-test. At the conclusion of the three study-test cycles, the participants then answered a question about their own use of interleaving (1 = Never or rarely; 6 = Almost always): In your own studies, how often do you mix up your studying (vs. focusing on one thing for an extended period at a time)? In general, participants reported low to moderate use of interleaving (M = 3.14; SD = 1.38), with only 25 (14%) of participants responding with a 5 or 6 (see Supplemental Online Materials Figure S1 for the histogram). They also responded to three motivation-related belief scales: growth mindset, difficult-as-importance, and difficulty-as-impossibility (see the Supplemental Online Materials Table S1 for the full list of these belief items and the descriptive statistics).

Results and Discussion

Replication of the interleaving effect. First, we replicated the now-robust interleaving effect. We conducted a linear mixed effects regression, predicting classification accuracy from schedule. Individual ID and categories nested within stimuli sets were included as crossed random effects. Interleaving (M = 0.62, SD = 0.28) led to significantly better classification test performance than did blocking (M = 0.51, SD = 0.24), b = 0.11, SE = 0.01, $p < .001$.

Next, to examine whether the interleaving effect varied across stimuli, we conducted a 2 (schedule) × 3 (stimuli set) ANOVA. There was no schedule × stimuli interaction, $F(2, 346) = 2.14$, $MSE = 0.06$, $p = .119$, $\eta_p^2 = .01$.[3] The marginal means and the Cohen's d effect size of the difference between the blocked and interleaved conditions for each stimuli set are described in Table 5.

Variation within individuals. As shown in Table 5, on average, participants' interleaved classification score was 11 percentage points higher than their blocked classification score. There was substantial variation within participants however—the standard deviation of the difference score was 16 percentage points. The bottom panel of Figure 2 shows the distribution of individuals' interleaved and blocked score differences (averaged across stimulus types) in Experiment 3. Although the majority of participants (n = 124 or 70%) displayed an overall interleaving benefit, a substantial minority of participants (30%) did not display an interleaving benefit. Note that for each participant, categories are randomly assigned to blocked and interleaved conditions; it is unclear whether this 30% represents people for whom interleaving is not better, or whether this reflects natural variations in item difficulty (e.g., easier categories being assigned to the blocked condition). Nevertheless, it raises the possibility that there is a subset of participants for whom interleaving is not better.

Our central research question was to examine how consistent the schedule benefits were within-subjects. Each participant engaged in three study-test cycles, each cycle with a different stimulus type. We examined within-participant consistency in two ways. First, we did a simple frequency count of how many times out of the three categories that participants experienced a blocking benefit. There was substantial variability. Out of 174 participants, 77 (44%) never experienced a blocking benefit, 61 (35%) experienced a blocking benefit on just one of the stimulus types, 31 (18%) experienced a blocking benefit on two of the stimulus types, and only five (3%) participants experienced blocking benefit on all three stimulus types. Conversely, only 15 (9%) never experienced an interleaving benefit, 49 (28%) experienced the interleaving benefit on just one of the stimulus types, 67 (39%) experienced an interleaving benefit on two stimulus types, and 43 (25%) experienced an interleaving benefit on all three stimulus types. These results sustain the possibility that there might be a small subset of participants who may benefit from blocked study over interleaved study.

Hence, the second way in which we examined within-participant consistency was to examine whether experiencing the size of a blocked or interleaved benefit on one set of stimulus types was related to the size of a blocked or interleaved benefit on either of the other two stimulus types. To test this,

[3] As expected, there was a main effect of schedule, $F(2, 173) = 93.40$, $MSE = 3.41$, $p < .001$, $\eta_p^2 = .35$. There was also a main effect of stimuli type, $F(2, 346) = 7.73$, $MSE = 0.42$, $p = .001$, $\eta_p^2 = .04$: classification performance of the butterflies ($M = 0.61$, $SD = 0.32$) was higher than classification performance of the artists ($M = 0.55$, $SD = 0.28$, $t(173) = 4.48$, $p < .001$, $d = 0.32$) or the fish ($M = 0.56$, SD = 0.33, $t(173) = 2.73$, $p = .007$, $d = 0.20$); classification performance of the artists and the fish were not significantly different from each other, $t(173) = 0.66$, $p = .52$.

we first calculated for each individual, the size of the interleaving benefit for each stimuli type by calculating the difference between their score on the interleaved categories and their score on the blocked categories. We then examined the correlation of the interleaving benefit between each pair of stimuli types. Neither the benefit in the artists stimuli nor the benefit in the butterflies stimuli were related to the benefit in the fish stimuli, $r(172) = -.01$, $p = .895$ and $r(172) = .09$, $p = .250$, respectively. There was a very weak correlation in interleaving benefit between the artists and butterflies stimuli, $r(172) = .19$, $p = .014$. Though significant, this correlation represents that only 3.6% of the variance in interleaving benefit between the two stimuli sets is shared. Moreover, a comparison of the three correlation coefficients (Lenhard & Lenhard, 2014) reveals that they are not statistically different from each other, $z = 1.03$, $p = .15$. We interpret this as evidence that there is unlikely to be a subset of participants for whom blocking is consistently better.

Overall, we see a robust interleaving effect across each of the three stimuli types, across categories within each of the three stimuli types, and within-individuals. Although there were a small handful of participants who showed an overall blocking benefit, further analysis showed that whether one benefited from blocking in one stimulus set was generally unrelated to what would be a better schedule for a different stimulus set.

Other sources of variation. We also examined other sources of variation. In terms of individual differences, we also collected participant responses to three motivation-related belief scales (growth mindset, difficulty-as-impossibility, difficulty-as-importance) as well as information about whether participants use interleaving in their own study. We examined whether any of these variables interacted with schedule, as that would indicate the benefit of interleaving varies across individuals on these variables. None of the variables significantly interacted with schedule, ps $> .37$; with one exception (difficulty-as-importance, $p = .046$). These analyses are reported in the Supplemental Online Materials (Tables S2 and S3).

We also examined variation from the studied categories themselves, which may differ in terms of difficulty or distinctiveness. Figure 5 illustrates the average classification test performance following blocked and interleaved study for each of the 24 categories and reveals the impressive replicability of the interleaving effect. The motor skills literature on interleaving has shown that more difficult skills often do not show an interleaving benefit (Wulf & Shea, 2002). To examine whether the interleaving benefit was moderated by overall category difficulty, we conducted a linear mixed effect model predicting classification accuracy from schedule, overall category difficulty (operationalized as the likelihood of correct classification across all participants and both schedule conditions) and the interaction between the two variables, allowing individual participant's intercepts to vary. Again, there was a robust interleaving benefit, $b = 0.11$, $SE = 0.01$, $t(3999) = 11.34$, $p < .001$ and no interaction, $b = 0.01$, $SE = 0.01$, $t(4020) = 0.65$, $p = .52$.

General Discussion

Prior studies have typically examined the average effects of interleaving, ignoring the potential for heterogeneity. Across all three experiments, we were able to repeatedly replicate the interleaving benefit across different stimuli types. These results are consistent with those reported by previous studies. Unlike prior studies, we delved beyond average effects, examining several potential sources of heterogeneity, from the task demands (categories of varying difficulty, working memory task demands) and from the participants themselves (reliability across multiple study-test cycles, working memory capacities).

In Experiments 1 and 2, the combination of individual WM load and manipulation of task demands on WM, we obtained a

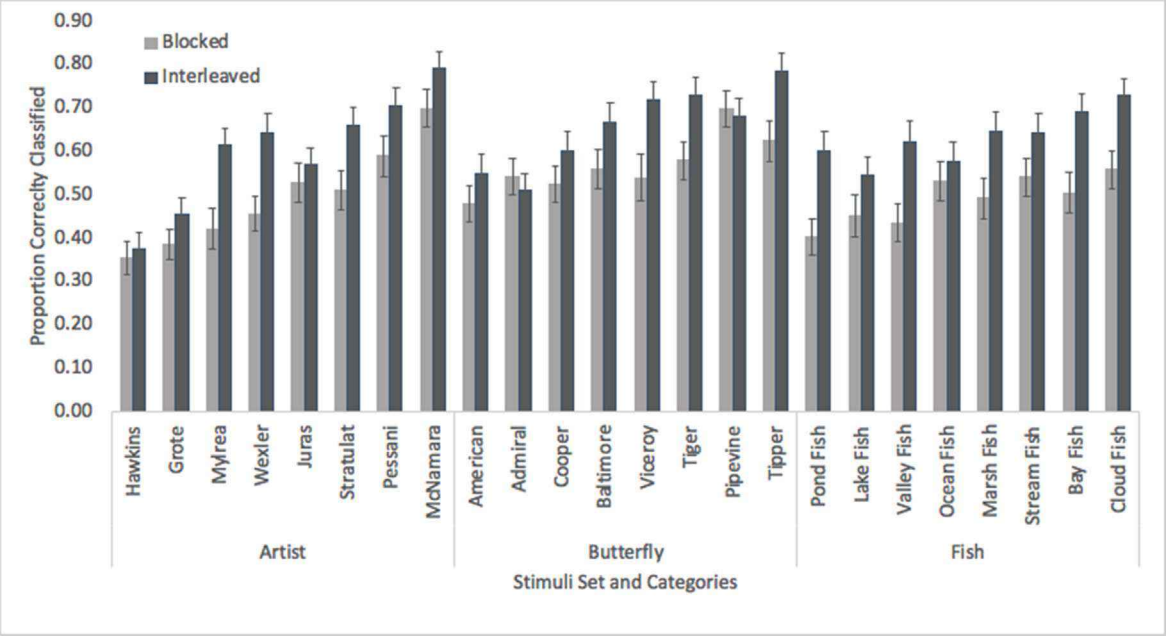

Figure 5. Classification performance by schedule and by category in Experiment 3. Error bars represent one standard error of the mean.

broad range of final test performance accuracy, ranging from floor to near-ceiling. These numbers indicate that we likely captured (nearly) the full spectrum of working memory load (either from the task demand or from capacity of the individual), and results show that across almost this entire spectrum, interleaved study yields better performance than blocked study.

In Experiment 3, we showed that although for any given study-test cycle, there is typically a subset of participants for whom interleaving is not better than blocking, participants who experience a blocked benefit under one stimuli type do not reliably experience a blocked benefit under the other stimuli types. Rather, heterogeneity may be more related to random noise than to anything intrinsic to the individual. Moreover, the design of Experiment 3—eight categories in each of three stimuli sets, for a total of 24 stimuli—allowed us to examine the robustness of the interleaving effect across categories of varying difficulty. We did not manipulate difficulty, but rather relied on the natural variations from the stimuli. We found that the interleaving benefit was not moderated by the difficulty of the categories—there was a remarkably consistent interleaving benefit across the 24 categories (eight per stimuli type).

These dimensions, of course, do not represent an exhaustive list of ways in which to examine the robustness of the interleaving effect. For example, in the current paper we focused only on visual categories. Brunmair and Richter (2019) in their meta-analysis reported finding that benefit of interleaving was more robust for visual categories but weaker for expository texts; in part, this may be a result of the visual categories being relatively "process-pure" in terms of category learning. That is, category-learning with visual categories is less likely to be affected by reading ability (e.g., comprehension, phonetic awareness, reading fluency) or to rely on multiple processes (e.g., solving mathematics problems involves not only being able to categorize a problem, but to accurately recall and implement the procedure or formula). While Brunmair and Richter have identified stimuli type as a boundary condition, more research is needed to understand how the underlying processes give rise to these heterogeneous effects.

Other studies have examined other dimensions of robustness too, such as robustness of the interleaving benefit after a delay (e.g., Carvalho & Goldstone, 2014b; Zulkiply & Burt, 2013). And while the majority of studies are conducted on young adults, other studies have shown that the interleaving benefit is robust across different ages, from elementary children learning handwriting skills (6 years of age; Ste-Marie et al., 2004) and mathematics (8–10 years of age; Nemeth et al., 2019) to older adults learning artists' painting styles (average age = 77 years; Kornell et al., 2010). In our present studies, we collected data only from university students; many studies nowadays are collected using online participant pools (e.g., Amazon Mechanical Turk, Prolific Academic) and we would recommend examining age as a moderator. Prior studies have shown that older adults have poorer associative memory performance compared to younger adults (e.g., Naveh-Benjamin, 2000); to the extent that the category learning requires associative memory (e.g., associating the relevant features with the correct category label), age could moderate interleaving effects.

Across the three experiments, another robust effect was that participants tended not to appreciate the benefits of interleaving. In Experiments 1 and 2, the majority of participants reported believing that blocking was better for their own learning even after most of them had just experienced an interleaving benefit; in Experiment 3, participants generally reported low to moderate use of interleaving in their own studying. Indeed, other research demonstrates that it is likely not just learners (see also McCabe, 2011; Yan et al., 2016) who fall prey to this metacognitive illusion that blocking is better than interleaving. In a survey of over 600 teachers, Yan and Oyserman (2017, poster) show that only a slight majority 59% of teachers thought that their students would learn more from interleaved practice than from blocked practice. In their examination of six popularly adopted seventh-grade math textbooks, Rohrer and colleagues (2020) found that only 9.7% of the 13,505 practice problems in the textbook were interleaved. In other words, despite research on interleaving, educational practices have not changed; the standard format in assignments and textbooks remains blocked. One possible reason for hesitancy of adopting interleaved practice methods, at least on the part of the educator, is the lack of sufficient research on the generalizability of the interleaving effect for all their students. We provide empirical evidence that demonstrates the benefit of interleaved practice within and across learners.

Conflict of Interest

The authors declare that they have no conflict of interest.

Acknowledgements

We thank Derek Stoeckenius, Melissa Walman, and Jingqi Yu for their help in collecting data, and the members of CogFog for their insightful feedback. Aspects of this research were presented at the 57th annual meeting of the Psychonomic Society, Boston, MA.

Author Contributions

V.X.Y. and F.S. conceptualized and designed Experiments 1 and 2. V.X.Y. conceptualized and designed Experiment 3. Data were collected with the help of research assistants under the supervision of V.X.Y. Data were analyzed by V.X.Y. The manuscript was drafted by both V.X.Y. and F.S.

Online Supplement

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jarmac.2021.05.002.

References

Abushanab, B., & Bishara, A. J. (2013). Memory and metacognition for piano melodies: Illusory advantages of fixed- over random-order practice. *Memory & Cognition,, 41*, 928–937. https://doi.org/10.3758/s13421-013-0311-z.

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*, 392–402. https://doi.org/10.3758/s13421-012-0272-7.

Brady, F. (2016). Contextual interference and teaching golf skills: Perceptual and motor skills. https://doi.org/10.2466/pms.1997.84.1.347

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

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

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

Carvalho, P. F., & Goldstone, R. L. (2017). March 23. The sequence of study changes what information is attended to, encoded, and remembered during category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 42*, 481–495. https://doi.org/10.1037/xlm0000406.

Conway, A. R. A., Kane, M. J., Bunting, M. F., Hambrick, D. Z., Wilhelm, O., & Engle, R. W. (2005). Working memory span tasks: A methodological review and user's guide. *Psychonomic Bulletin & Review, 12*, 769–786. https://doi.org/10.3758/bf03196772.

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

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*, 1088–1101. https://doi.org/10.3758/s13421-019-00918-4.

Goode, S., & Magill, R. A. (1986). Contextual interference effects in learning three badminton serves. *Research Quarterly for Exercise and Sport, 57*, 308–314. https://doi.org/10.1080/02701367.1986.10608091.

Guadagnoli, M. A., Holcomb, W. R., & Weber, T. (1999). The relationship between contextual interference effects and performer expertise on the learning of a putting task. *Journal of Human Movement Studies, 37*, 19–36.

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

Hall, K. G., Domingues, D. A., & Cavazos, R. (1994). Contextual interference effects with skilled baseball players. *Perceptual and Motor Skills, 78*, 835–841. https://doi.org/10.2466/pms.1994.78.3.835.

Hebert, E. P., Landin, D., & Solmon, M. A. (1996). Practice schedule effects on the performance and learning of low-and high-skilled students: An applied study. *Research Quarterly for Exercise and Sport, 67*, 52–58. https://doi.org/10.1080/02701367.1996.10607925.

Jones, L. L., & French, K. E. (2007). Effects of contextual interference on acquisition and retention of three volleyball skills. *Perceptual and Motor Skills, 105*, 883–890. https://doi.org/10.2466/pms.105.3.883-890.

Kane, M. J., Hambrick, D. Z., Tuholski, S. W., Wilhelm, O., Payne, T. W., & Engle, R. W. (2004). The generality of working memory capacity: A latent-variable approach to verbal and visuospatial memory span and reasoning. *Journal of Experimental Psychology. General, 133*, 189–217. https://doi.org/10.1037/0096-3445.133.2.189.

Kang, S. H. K., & Pashler, H. (2012). Learning painting styles: Spacing is advantageous when it promotes discriminative contrast. *Applied Cognitive Psychology, 26*, 97–103. https://doi.org/10.1002/acp.v26.110.1002/acp.1801.

Kirk-Johnson, A., Galla, B. M., & Fraundorf, S. H. (2019). Perceiving effort as poor learning: The misinterpreted-effort hypothesis of how experienced effort and perceived learning relate to study strategy choice. *Cognitive Psychology, 115*, 1–31. https://doi.org/10.1016/j.cogpsych.2019.101237.

Kornell, N., & Bjork, R. A. (2008). Learning concepts and categories: Is spacing the "enemy of induction"?. *Psychological Science, 19*, 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.

Lenhard, W., & Lenhard, A. (2014). Hypothesis tests for comparing correlations. Bibergau (Germany): Psychometrica. https://doi.org/10. 13140/RG.2.1.2954.1267

Lin, C. H. J., Chiang, M. C., Wu, A. D., Iacoboni, M., Udompholkul, P., Yazdanshenas, O., & Knowlton, B. J. (2012). Age related differences in the neural substrates of motor sequence learning after interleaved and repetitive practice. *Neuroimage, 62*, 2007–2020. https://doi.org/10.1016/j.neuroimage.2012.05.015.

Lin, C.-H., Knowlton, B. J., Wu, A. D., Iacoboni, M., Yang, H.-C., Ye, Y.-L., & ... Chiang, M.-C. (2016). Benefit of interleaved practice of motor skills is associated with changes in functional brain network topology that differ between younger and older adults. *Neurobiology of Aging, 42*, 189–198. https://doi.org/10.1016/j.neurobiolaging.2016.03.010.

Magill, R. A., & Hall, K. G. (1990). A review of the contextual interference effect in motor skill acquisition. *Human Movement Science, 9*, 241–289.

McCabe, J. (2011). Metacognitive awareness of learning strategies in undergraduates. *Memory & Cognition, 39*, 462–476. https://doi.org/10.3758/s13421-010-0035-2.

McDaniel, M. A., Cahill, M. J., Robbins, M., & Wiener, C. (2014). Individual differences in learning and transfer: Stable tendencies for learning exemplars versus abstracting rules. *Journal of Experimental Psychology: General, 143*, 668–693. https://doi.org/10.1037/a0032963.

Naveh-Benjamin, M. (2000). Adult age differences in memory performance: Tests of an associative deficit hypothesis. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 26*, 1170–1187.

Nemeth, L., Werker, K., Arend, J., Vogel, S., & Lipowsky, F. (2019). Interleaved learning in elementary school mathematics: Effects on the flexible and adaptive use of subtraction strategies. *Frontiers in Psychology, 10*. https://doi.org/10.3389/fpsyg.2019.00086.

Ollis, S., Button, C., & Fairweather, M. (2005). The influence of professional expertise and task complexity upon the potency of the contextual interference effect. *Acta Psychologica, 118*, 229–244. https://doi.org/10.1016/j.actpsy.2004.08.003.

Porter, J. M., & Magill, R. A. (2010). Systematically increasing contextual interference is beneficial for learning sport skills. *Journal of Sports Sciences, 28*, 1277–1285.

Rohrer, D. (2012). Interleaving helps students distinguish among similar concepts. *Educational Psychology Review, 24*, 355–367. https://doi.org/10.1007/s10648-012-9201-3.

Rohrer, D., Dedrick, R. F., & Hartwig, M. K. (2020). The scarcity of interleaved practice in mathematics textbooks. *Educational Psychology Review, 32*, 873–883. https://doi.org/10.1007/s10648-020-09516-2.

Rozenshtein, A., Pearson, G. D. N., Yan, S. X., Liu, A. Z., & Toy, D. (2016). Effect of massed versus interleaved teaching method on performance of students in radiology. *Journal of the American College of Radiology, 13*, 979–984. https://doi.org/10.1016/j.jacr.2016.03.031.

Sana, F., Yan, V. X., & Kim, J. A. (2017). Study sequence matters for the inductive learning of cognitive concepts. *Journal of Educational Psychology, 109*, 84–98. https://doi.org/10.1037/edu0000119.

Sana, F., Yan, V. X., Kim, J. A., Bjork, E. L., & Bjork, R. A. (2018). Does working memory capacity moderate the interleaving benefit?. *Journal of Applied Research in Memory and Cognition, 7*, 361–369. https://doi.org/10.1016/j.jarmac.2018.05.005.

Shea, C. H., Kohl, R., & Indermill, C. (1990). Contextual interference: Contributions of practice. *Acta Psychologica, 73*, 145–157. https://doi.org/10.1016/0001-6918(90)90076-R.

Simon, D. A., & Bjork, R. A. (2001). Metacognition in motor learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 27*, 907–912. https://doi.org/10.1037/0278-7393.27.4.907.

Ste-Marie, D. M., Clark, S. E., Findlay, L. C., & Latimer, A. E. (2004). High levels of contextual interference enhance handwriting skill acquisition. *Journal of Motor Behavior, 36*, 115–126. https://doi.org/10.3200/JMBR.36.1.115-126.

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

Turner, M. L., & Engle, R. W. (1989). Is working memory capacity task dependent?. *Journal of Memory and Language, 28*, 127–154. https://doi.org/10.1016/0749-596X(89)90040-5.

Unsworth, N., & Engle, R. W. (2007). The nature of individual differences in working memory capacity: Active maintenance in primary memory and controlled search from secondary memory. *Psychological Review, 114*, 104–132. https://doi.org/10.1037/0033-295X.114.1.104.

Unsworth, N., Heitz, R. P., Schrock, J. C., & Engle, R. W. (2005). An automated version of the operation span task. *Behavior Research Methods, 37*, 498–505. https://doi.org/10.3758/bf03192720.

Vlach, H. A., Sandhofer, C. M., & Kornell, N. (2008). The spacing effect in children's memory and category induction. *Cognition, 109*, 163–167. https://doi.org/10.1016/j.cognition.2008.07.013.

Wang, J., Liu, Z., Xing, Q., & Seger, C. A. (2020). The benefit of interleaved presentation in category learning is independent of working memory. *Memory, 28*, 285–292. https://doi.org/10.1080/09658211.2019.1705490.

Wulf, G., & Shea, C. H. (2002). Principles derived from the study of simple skills do not generalize to complex skill learning. *Psychonomic Bulletin & Review, 9*, 185–211.

Yan, V. (2014). Learning concepts and categories from examples: How learners' beliefs match and mismatch the empirical evidence. UCLA. ProQuest ID: Yan_ucla_0031D_12577. Merritt ID: ark:/13030/m5mk7t7j. Retrieved from https://escholarship.org/uc/item/91q7z7z4

Yan, V. X., Garcia, M. A., Bjork, R. A., & Bjork, E. L. (2014, May). Best of Both Worlds? Combining Blocked and Interleaved Schedules in Category Learning. Poster presented at 26th Annual Convention of the Association for Psychological Science, San Francisco, CA

Yan, V. X. & Oyserman, D. (2017, Nov). Linking mindsets with toolsets: Interpretations of experienced difficulty matter for knowing how to learn. Poster presented at the 58th Annual Scientific Meeting of the Psychonomic Society, Vancouver, Canada.

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*, 918–933. https://doi.org/10.1037/xge0000177.

Zulkiply, N., & Burt, J. S. (2013). Inductive learning: Does interleaving exemplars affect long-term retention?. *Malaysian Journal of Learning and Instruction, 10*, 133–155.

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*, 215–221. https://doi.org/10.1016/j.learninstruc.2011.11.002.

Received July 21, 2020
Received in revised form May 10, 2021
Accepted May 10, 2021
Available Online: 22 June 2021