Why does interleaving improve math learning? The contributions of discriminative contrast and distributed practice

Nathaniel L. Foster¹ · Michael L. Mueller² · Christopher Was² · Katherine A. Rawson² · John Dunlosky²

Published online: 15 March 2019
© The Psychonomic Society, Inc. 2019

Abstract

Interleaved practice involves studying exemplars from different categories in a non-systematic, pseudorandom order under the constraint that no two exemplars from the same category are presented consecutively. Interleaved practice of materials has been shown to enhance test performance compared to blocked practice in which exemplars from the same category are studied together. Why does interleaved practice produce this benefit? We evaluated two non-mutually exclusive hypotheses, the *discriminative-contrast hypothesis* and the *distributed-practice hypothesis*, by testing participants' performance on calculating the volume of three-dimensional geometric shapes. In Experiment 1, participants repeatedly practiced calculating the volume of four different-sized shapes according to blocked practice, interleaved practice, or remote-interleaved practice (which involved alternating the practice of volume calculation with non-volume problems, like permutations and fraction addition). Standard interleaving enhanced performance compared to blocked practice but did not produce enhanced performance compared to remote interleaving. In Experiment 2, we replicated this pattern and extended the results to include a remote-blocked group, which involved blocking volume calculation with non-volume problems. Performance on key measures was better for remote-interleaved groups compared to remote-blocked groups, a finding that supports the distributed-practice hypothesis.

Keywords Interleaved practice · Distributed practice effect · Math learning · Practice schedules

Introduction

Interleaving is a study technique that involves studying exemplars from different categories in a non-systematic, pseudorandom order under the constraint that no two exemplars from the same category are presented consecutively. For example, *interleaved* practice of math concepts might involve intermixing problems involving fraction addition with fraction subtraction, whereas *blocked* practice would involve practicing a block of fraction addition problems followed by a block of fraction subtraction problems. Interleaving practice of to-be-learned information is a promising strategy for improving learning of various kinds of material (for reviews, see Dunlosky, Rawson, Marsh, Nathan, & Willingham, 2013; Rohrer, 2012). Relative to blocked practice, interleaving has

been shown to improve learning of different mathematics materials, including permutation calculations (Rohrer & Taylor, 2007; Taylor & Rohrer, 2010), identifying the number of faces, corners, edges, and angles of prism shapes (Taylor & Rohrer, 2010), calculating the volume of geometric shapes using task-relevant formulas (Rohrer, 2012; Rohrer & Taylor, 2007), and applying statistical tests to different word problems (Sana, Yan, & Kim, 2017).

Despite growing evidence that interleaving is beneficial, not much experimental research has focused on attempting to reveal why interleaving works. The primary goal for the current study was to evaluate explanations for why interleaved practice enhances performance on criterion tests, which we refer to as *interleaving effects*.

Theoretical accounts for interleaving effects

Interleaving effects can be explained by two non-mutually exclusive hypotheses. According to the discriminative-contrast hypothesis, interleaved practice improves learning of a problem type because instances of that problem type are practiced in close proximity to instances of different problem types from a similar domain. When this happens, people are

✉ Nathaniel L. Foster
nlfoster@smcm.edu

¹ Department of Psychology, St. Mary's College of Maryland, 47645 College Dr., St. Mary's City, MD 20686, USA

² Kent State University, 800 East Summit St., Kent, OH 44240, USA

more likely to notice similarities and differences between problem-type solutions, which can improve performance on later criterion tests (Birnbaum, Kornell, Bjork, & Bjork, 2013; Carvalho & Goldstone, 2014, 2019; Kang & Pashler, 2012; Kornell & Bjork, 2008; Kornell, Castel, Eich, & Bjork, 2010; Zulkiply & Burt, 2013). For example, Kang and Pashler (2012) had participants study paintings from three painters according to one of three practice schedules. The blocked practice group studied all eight paintings from the first artist consecutively, followed by study of the eight paintings from the second artist and then the eight paintings from the third artist. The interleaved practice group studied paintings in eight blocks of three paintings, with each block containing one painting from each of the three artists. Finally, the blocked-with-cartoons group studied target paintings that were interleaved with cartoons rather than with each other, which limited participants' ability to directly contrast paintings from different artists. More specifically, for any given trial, participants studied a painting by an artist for 5 s, then they had 10.5 s to read a cartoon drawing, and then they studied another painting by the same artist. This sequence continued until all the paintings from one artist were presented, and then the sequence was repeated for the other two artists. The interleaved practice group outperformed both the blocked practice group and the blocked-with-cartoons group on a transfer test involving identification of new paintings by the same artists. These findings support the discriminative-contrast hypothesis because (a) the interleaved practice group was the only group that was presented with paintings from the different artists consecutively and hence could more easily contrast the styles of the different painters, and (b) spacing was controlled for in the blocked-with-cartoons group (Kang & Pashler, 2012; see also Mitchell, Nash, & Hall, 2008).

Further support for the discriminative-contrast hypothesis comes from research by Wahlheim, Dunlosky, and Jacoby (2011). Participants studied bird species according to a blocked or interleaved schedule. Critically, on each trial, participants studied exemplars one at a time (singles group) or in pairs (pairs group). In particular, participants in the singles group were presented with a bird exemplar on each trial, whereas participants in the pairs group were presented with two bird exemplars side by side. Successful categorization of novel exemplars was greater for interleaved than for blocked study and the effect was larger for the pairs group than for the singles group, suggesting that interleaving works best when participants can directly contrast different bird species.

Although prior evidence indicates that discriminative contrast is sufficient to produce an interleaving effect, it may not always be necessary to show the benefits of interleaving. For example, Rohrer, Dedrick, and Burgess (2014) showed that interleaving benefits learning in the classroom even when practice consists of consecutively dissimilar problems. Students in grade 7 learned to solve four types of math problems according

to either a blocked or an interleaved practice schedule. Importantly, unlike the highly similar problem types used in past research on math and interleaving (Rohrer & Taylor, 2007; Taylor & Rohrer, 2010), problem types used by Rohrer et al. (2014) were superficially dissimilar. The authors argued that if interleaving benefits learning only because it increases discrimination between problem types, the benefit should not occur for problems that are superficially dissimilar and thus easily discriminated. Inconsistent with this prediction, Rohrer et al. (2014) observed a benefit of interleaving practice of dissimilar problems on a delayed test (see also Rohrer, Dedrick, & Stershic, 2015, for another example of interleaving dissimilar problems). If the benefits of interleaving arise due to increased opportunities to discriminate different problem types that are from a similar problem domain, why did interleaved practice of *dissimilar* material produce enhanced learning?

An alternative to the discriminative-contrast hypothesis of interleaved practice is the *distributed-practice hypothesis*. Research investigating the effect of repetition on memory has indicated a substantial benefit of distributed repetition compared to massed repetition (Cepeda, Pashler, Vul, Wixted, & Rohrer, 2006). These benefits have been observed across different learning contexts, including verbal learning (e.g., Janiszewski, Noel, & Sawyer, 2003), procedural learning (e.g., Donovan & Radosevich, 1999), problem solving (e.g., Jacoby, 1978), and math learning in the classroom (e.g., Schutte et al., 2015). Rohrer et al. (2014) argued that distributing practice of math problems by way of an interleaved practice schedule strengthens the association between a problem and the correct solution more so than does blocked practice. This greater strengthening between a problem and its strategy for distributed than blocked practice may result from several theoretical mechanisms, such as by study-phase retrieval practice or by encoding variability (for reviews, see Benjamin & Tullis, 2010; Donovan & Radosevich, 1999; Hintzman, 1974; Hintzman, Summers, & Block, 1975). Most important here is that these outcomes and rationale indicate that distributing practice alone could produce an interleaving effect without the contribution of discriminative contrast.

Critical for our purposes is the issue of whether interleaved practice produces benefits simply because practice of target problems for each type is distributed across the study phase. If so, enhancement of target activity should occur regardless of the type of interpolated activity because the benefit of interleaving would arise from the strengthening of associations between problem types and the strategies needed to solve them (Rohrer et al., 2014).

Experiment 1

Experiment 1 was designed to evaluate the degree to which the benefits of interleaving arise due to discriminative

contrast, distributed practice, or a combination of both. To date, research on interleaving of math materials (e.g., Rohrer & Taylor, 2007; Taylor & Rohrer, 2010) has indicated robust benefits of interleaving compared to blocked practice. However, in most cases, interleaving conflated discriminative contrast and distributed practice, thus making it impossible to evaluate which factor (or factors) was producing the overall interleaving effect. More specifically, in standard interleaving procedures, any given problem type is both contrasted consecutively with a different problem type and practiced in a distributed fashion across the study phase. Testing whether interleaving of dissimilar materials benefits learning (as compared to blocked practice) is one way to help further understand the underlying mechanisms of interleaved practice because when the problem types are highly dissimilar then participants would not need to learn what aspects of each problem are relevant to assigning it to a problem type. That is, when all the types of problems are highly different, then discriminative contrast would likely not be needed to learn how to solve them. For instance, if one were learning to identify which artist produced a given painting and each artist had a very different style (e.g., only one was a cubist, another was a realist, and another an impressionist), then fine-grained discrimination would not necessarily be needed to learn which painter produced which painting (because each *kind* of painting is obviously distinct). Rohrer et al. (2014) did use dissimilar problem types (see above for discussion) and found an interleaving effect, which provides more support for the distributed-practice hypothesis than for the discriminative-contrast hypothesis. Even so, the results of Rohrer et al. (2014) do not rule out the possibility that discriminative contrast across similar problem types can further benefit learning, because only two groups were compared: interleaved versus blocked practice with both involving dissimilar materials. Whether discriminative contrast could further enhance the benefits of interleaving remains unknown.

To tease apart the contributions of discriminative contrast versus distributed practice in a laboratory setting, we had college students learn to solve math problems originally used in research by Rohrer and Taylor (2007). These problems required participants to calculate the volumes of different geometric shapes. Participants were randomly assigned to blocked, standard-interleaved, and remote-interleaved groups. In the blocked group, participants practiced calculating volumes of four instances for each of four different geometric shapes: wedge, spheroid, spherical cone, and half cone (cf. Rohrer & Taylor, 2007). In the standard-interleaved group, practice involved solving one instance of each of the four shapes, followed by three more blocks of the same order but with different instances in each block. Critically, in the remote-interleaved group, participants practiced wedge volume problems interleaved with non-volume problems like fraction addition, fraction division, and permutations. Doing

so would reduce engagement of a contrast process that promotes discrimination between the different types of volume problems. If discriminative contrast of similar materials (e.g., learning different kinds of volume problems) provides an added benefit for observing interleaving effects, final test performance for the wedge problems (which we refer to as the *critical problems* from this point forward) will be greater for the standard-interleaved group as compared to the remote-interleaved group. Alternatively, if distributed practice alone produces interleaving benefits, the standard-interleaved and remote-interleaved groups will perform equally and will both outperform the blocked group. Discriminative contrast and distributed practice may combine during interleaved practice to enhance learning. If so, the standard-interleaved group will outperform the remote-interleaved group, which will outperform the blocked group.

Finally, we measured both overall accuracy of volume calculations as well as accuracy of formula retrieval, because participants may learn the association between a geometric shape and the formula for calculating the volume of that shape, but they may fail to produce the correct answer due to a computation error (such as substituting the radius value in for the height value). Thus, poor computation performance can decrease overall accuracy even if students have retrieved and are using the correct formula for a problem. Accordingly, we present analyses pertaining to how practice schedules affected both accuracy of formula retrieval alone as well as overall accuracy, which is composed of both formula retrieval and computation performance.

Method

Participants, design, and materials

Participants ($N = 126$) were recruited from Kent State University to fulfill a partial course requirement and were randomly assigned to one of three practice groups: blocked ($n = 43$), interleaved ($n = 40$), or remote interleaved ($n = 43$). A power analysis conducted using G*Power 3.1.9.2 (Faul, Erdfelder, Lang, & Buchner, 2007) for a one-way ANOVA with power set at .80, $\alpha = .05$, and number of groups set to three indicated that this sample size afforded sufficient sensitivity to detect medium or larger effect sizes ($f > .25$). We set power analysis to detect medium or larger than medium effect sizes because (1) the effect sizes reported in Rohrer and Taylor (2007) were medium to large, and (2) effects of less than medium would not be of much practical significance. In the blocked and interleaved practice groups, participants learned to compute the volume of four three-dimensional geometric shapes (i.e., wedges, spheroids, spherical cones, and half cones) taken from Rohrer and Taylor (2007). The shapes and their respective formulas are shown in Fig. 1. In the remote-interleaved practice group, participants learned to

Fig. 1 Geometric shapes and their formulas as used in Rohrer and Taylor (2007). Reprinted from Rohrer and Taylor (2007)

compute volumes for wedges, add fractions (e.g., $\frac{1}{2} + \frac{3}{8} = ?$), divide exponents (e.g., $x^3 \div x^2 = ?$), and solve permutation problems (e.g., how many ways can the letters aaabbbcc be arranged?).

Procedure

Participants completed three separate phases: pre-practice test (henceforth *pretest* for brevity), practice, and final test. The pretest contained ten problems in total: four volume problems (one for each of the four geometric shapes), two fraction add,

two exponent division, and two permutation problems. Participants were given up to 4 min to complete the pretest. This pretest was identical across groups. Upon completion of the pretest, participants then proceeded to the practice phase.

In the practice phase, participants were seated at a computer that presented a timed slideshow. A schematic representation of the study order for each practice group is presented in Table 1. The order of presentation of practice problems was the same for all participants within each practice group. In each group participants were given a tutorial on how to solve the type of problem for 60 s each (see Fig. 2a). During the

Table 1 Practice orders for each group

Practice group	Practice order
Blocked	$T_{VW} VW_1 VW_2 VW_3 VW_4 T_{VS} VS_1 VS_2 VS_3 VS_4 T_{VC} VC_1 VC_2 VC_3 VC_4 T_{VH} VH_1 VH_2 VH_3 VH_4$
Interleaved	$T_{VW} T_{VS} T_{VC} T_{VH} VW_1 VS_1 VC_1 VH_1 VW_2 VS_2 VC_2 VH_2 VW_3 VS_3 VC_3 VH_3 VW_4 VS_4 VC_4 VH_4$
Remote interleaved	$T_{VW} T_{ED} T_{FA} T_P VW_1 ED_1 FA_1 P_1 VW_2 ED_2 FA_2 P_2 VW_3 ED_3 FA_3 P_3 VW_4 ED_4 FA_4 P_4$
Remote blocked	$T_{VW} VW_1 VW_2 VW_3 VW_4 T_{ED} ED_1 ED_2 ED_3 ED_4 T_{FA} FA_1 FA_2 FA_3 FA_4 T_P P_1 P_2 P_3 P_4$

T Tutorial, *VW* Volume Wedge, *VS* Volume Spheroid, *VC* Volume Spherical Cone, *VH* Volume Half Cone, *ED* Exponent Divide, *FA* Fraction Add, *P* Permutation

Numbers 1–4 indicate a unique instance of each problem type

The Remote Blocked practice group was only included in Experiment 2. For half of the participants in Experiment 2, the orders of wedge and spheroid problems were switched in the blocked and interleaved groups, and wedge problems were replaced with spheroid problems in the remote groups

a**Lesson for: Volume: WEDGE**

How do you solve this problem?

A wedge is the boldfaced portion of the tube shown below. Its bottom is a circle, and its top is a slanted oval.

When solving volume problems you will be given the radius (r) and the height (h).

Its volume equals:

$$\frac{r^2 h \pi}{2}$$

For example, if you were told to solve for the volume of a wedge with $r = 2$ and $h = 4$, then you would do the following steps.

First write the formula for the wedge next to the problem number in the box:

$$\boxed{\frac{r^2 h \pi}{2}}$$

Next plug the radius and height into the formula (writing the radius twice because it is squared):

$$\frac{\cancel{2} \cdot 2 \cdot 4\pi}{\cancel{2}}$$

Simplify your answer by crossing out numbers in the numerator and denominator that match. Solve by multiplying and dividing the remaining numbers. Write the answer in the oval on the answer sheet: $\textcircled{8\pi}$

Fig. 2 **a** An example of a tutorial trial during the practice phase. Shape graphics are reprinted from Rohrer and Taylor (2007). **b** An example of a wedge trial and subsequent feedback during the practice phase

tutorials, participants were instructed to not write down the formula while studying the lessons. Participants were given four practice problems of each type to solve for 30 s each (see Fig. 2b). Participants recorded their responses to each practice problem in an answer booklet that had one response on each page to discourage them from looking back at previous responses. Participants were instructed to recall the formula for each volume problem and then compute the answer. For all other problem types, participants were instructed to simply write the answer. After attempting to solve each problem, the computer displayed the correct solution for 15 s (see Fig. 2b).

Participants returned 1 week later for the final test. The test was presented in a slideshow individually on a computer. Four novel instances of each problem type were presented, and participants had up to 45 s to solve each one. Similar to practice, participants were instructed to respond with both the formula and the computed value for the volume problems. Importantly, the final test output order was controlled to have participants in all groups attempt to solve all of the volume wedge problems first. All groups attempted to solve the remaining problem types in a blocked order that corresponded to the order in which they were studied. Thus, the blocked and interleaved groups were tested in the following order: wedges,

b

Please solve the following problem and write your answer on space number 1 on the answer sheet provided.

1.)

Find the volume of a wedge with $r = 1$ and $h = 2$

Write the Formula in the

And the answer in the

This is a VOLUME: WEDGE problem

FEEDBACK:

The solution to number 1 is:

$$1\pi$$

Remember the formula for VOLUME: WEDGE problems is

$$\frac{r^2 h \pi}{2}$$

Thus, plugging in $r = 1$ and $h = 2$ to the formula gives

$$\frac{1 \cdot 1 \cdot 2\pi}{2}$$

Fig. 2 continued.

spheroids, spherical cones, half cones. The remote-interleaved group was tested in the following order: wedges, exponent division, fraction addition, and permutation.¹

Results and discussion

In this section, we first present participants' pre-knowledge of math materials by analyzing their accuracy on the pretest as well as performance during the practice phase. We then report our analysis of participants' accuracy of formula retrieval on

the final test, first for wedges and then for all four geometric shapes. Finally, we present final test performance for the wedges (the critical problems) and then for all four shapes. To evaluate the distributed-practice hypothesis, we compared the blocked group to both the remote-interleaved and standard-interleaved groups. To evaluate the discriminative-contrast hypothesis, we compared the standard-interleaved group to the remote-interleaved group.

Pretest performance and practice performance

A one-way ANOVA on performance of all problems on the pretest indicated no significant difference between the blocked group ($M = .31$, $SD = .12$), the interleaved group

¹ We administered a second test that included all problems from the practice phase. We do not report results from this test because they largely mirrored the results of the initial final test.

($M = .34$, $SD = .12$), and the remote-interleaved group ($M = .32$, $SD = .12$), $F(2,123) = .48$, $MSE = .015$, $p = .62$, $\eta^2 = .007$. Pretest scores were not significantly different from each other when the analysis was restricted to the geometric shape problems for the blocked group ($M = .07$, $SD = .11$), the interleaved group ($M = .08$, $SD = .13$), and the remote-interleaved group ($M = .06$, $SD = .12$), $F(2,123) = .21$, $MSE = .015$, $p = .81$, $\eta^2 = .003$.

We analyzed performance on the practice phase for the critical problems to explore whether we replicated past research showing better practice performance in the blocked group compared to the interleaved group (Rohrer & Taylor, 2007). Practice performance for critical problems was not significantly different for the blocked group ($M = .90$, $SD = .23$) compared to the interleaved group ($M = .88$, $SD = .22$), $t(81) = .41$, $p = .68$, $d = .09$. Practice performance for critical problems was also not significantly different for the blocked group compared to the remote-interleaved group ($M = .87$, $SD = .25$), $t(84) = .57$, $p = .57$, $d = .12$. Finally, practice performance for critical problems was not significantly different for the interleaved group compared to the remote-interleaved group, $t(81) = .17$, $p = .87$, $d = .04$. Note, however, to provide the closest replication of Rohrer and Taylor (2007), we also examined practice performance for all volume problems for the blocked group ($M = .90$, $SD = .17$) and the interleaved group ($M = .75$, $SD = .25$), $t(81) = 3.32$, $p = .001$, $d = .73$, which did demonstrate the typical lower practice performance for the interleaved group.

Formula retrieval

Responses were counted as correct if a participant wrote the correct volume formula in their booklets. Of primary interest, our first set of analyses focused on formula retrieval accuracy for the wedge volume problems, the critical problem type studied (and tested first) by all groups.

The mean proportion of correct formula retrieval for wedge problems is plotted in Fig. 3. A one-way ANOVA on formula retrieval for wedge problems indicated no significant effect of group, $F(2,123) = 1.82$, $MSE = .232$, $p = .17$, $\eta^2 = .03$.

Accuracy of formula retrieval for all geometric shapes on the final test is presented in the top row of Table 2. A one-way ANOVA indicated a significant effect of group on formula retrieval for all geometric shapes on the final test, $F(2,123) = 11.07$, $MSE = .147$, $p < .001$, $\eta^2 = .15$. Formula retrieval was significantly greater for the interleaved group than the blocked group, $t(81) = 2.83$, $p < .01$, $d = .62$. Formula retrieval was also significantly greater for the remote-interleaved group than the interleaved group, $t(81) = 1.99$, $p = .05$, $d = .44$, and the blocked group, $t(84) = 4.68$, $p < .001$, $d = 1.00$.

Fig. 3 Mean proportion of correct formula retrieval for wedge problems during final test in Experiment 1. Error bars represent the standard error of the mean

Final test performance

The mean final test performance for wedges for each group is plotted in Fig. 4. A one-way ANOVA on final performance of wedge problems indicated a significant effect of group, $F(2,123) = 3.44$, $MSE = .192$, $p = .04$, $\eta^2 = .05$. Final test performance for wedges was not significantly greater for the interleaved group than the blocked group, $t(81) = .87$, $p = .39$, $d = .20$. Final test performance for wedges was not significantly greater for the remote-interleaved group compared to the interleaved group, $t(81) = 1.65$, $p = .10$, $d = .36$, but was significantly greater for remote-interleaved than the blocked groups, $t(84) = 2.59$, $p = .01$, $d = .56$.

Mean final test performance for all problems is presented in the second row of Table 2. A one-way ANOVA on final performance indicated a significant effect of group, $F(2,123) = 35.50$, $MSE = .073$, $p < .001$, $\eta^2 = .37$. Final test performance was significantly greater for the interleaved group than for the blocked group, $t(81) = 3.05$, $p = .003$, $d = .67$, for the remote-interleaved group than the interleaved group, $t(81) = 4.62$, $p < .001$, $d = 1.01$, and for the remote-interleaved group than the blocked group, $t(84) = 10.16$, $p < .001$, $d = 2.19$.

In summary, the interleaving effect – which compared the standard interleaving group to the blocked group – was shown to benefit both retrieval of formulas and final test performance when all problems were considered. When analysis was restricted to the wedge problems, the interleaving effect was not significant. Interestingly, standard interleaving alone did not beat remote interleaving on any of the dependent measures. If anything, the remote-interleaving group performed numerically better than the standard interleaving group on formula retrieval for wedges (although this effect did not reach significance, see Fig. 3).

Table 2 Mean formula retrieval and final test performance for all problems in Experiments 1 and 2

Experiment	Item type	Group			
		Blocked	Interleaved	Remote blocked	Remote interleaved
1	Formula retrieval	.20 (.04)	.40 (.06)	--	.59 ¹ (.07)
	Final test performance	.18 (.04)	.39 (.06)	--	.68 ² (.03)
2	Formula retrieval	.12 (.03)	.27 (.06)	.35 ³ (.07)	.54 ³ (.07)
	Final test performance	.12 (.02)	.24 (.04)	.58 ⁴ (.03)	.62 ⁴ (.03)

Standard error of the mean are presented in parentheses

¹ This value is based on wedge formula retrieval success

² This value is based on accuracy of wedge volume calculation, fraction addition, exponent division, and permutation problems

³ This value is based on retrieval success of wedge formula or spheroid formula

⁴ This value is based on accuracy of wedge or spheroid calculation, fraction addition, exponent division, and permutation problems

Experiment 2

The results of Experiment 1 suggest that standard interleaving does not produce superior formula retrieval on the critical (wedge) problems (in which the retention interval between practice and the final test was held constant for all groups) compared to remote interleaving. Instead, remote interleaving tended to support higher performance than did standard interleaving. These outcomes provide more support for the distributed-practice hypothesis than for the discriminative-contrast hypothesis, and even suggest that discriminative contrast may not contribute significantly to the interleaving effect in the present context.

One limitation to this conclusion pertains to a methodological difference between the blocked, interleaved, and remote-interleaved groups that could explain the enhanced performance for the latter group. In particular, participants in the remote-interleaved group may do well because they only need to learn how to solve one type of volume problem, namely wedge problems (whereas the interleaved and blocked groups

attempted to learn all the problem types). This explanation is based on the principle of cue overload wherein the effectiveness of a retrieval cue is reduced as more targets become associated with that cue (Watkins & Watkins, 1975). That is, when considering the cue of “volume,” participants in the blocked and interleaved groups may experience cue overload because four formulas are all associated with the concept of “volume.” But the remote-interleaved group associated the volume cue with a single formula, which should not produce cue overload.

To provide a more definitive evaluation of the two focal hypotheses, we need to compare the size of the interleaving effect when cue-overload may influence performance to its size when cue-overload cannot influence performance. To do so, in Experiment 2, we included a fourth group of participants who practiced volume problems of one type as the first four problems during practice, and then practiced four fraction addition problems, four exponent division problems, and four permutations problems. This *remote-blocked* practice group served as a comparison group to the remote-interleaved group (both of which would be minimally influenced by cue-overload) to evaluate whether the benefits of interleaving arise due to discriminative contrast and/or distributed practice. The predictions are as follows. If discriminative contrast contributes to the present interleaving effect, then the size of the interleaving effect will be larger for the standard-interleaved versus the standard-blocked group than for the remote-interleaved versus the remote-blocked group. In contrast, if the interleaving effect arises from distributed practice alone, then the interleaving effect will be the same size for the standard groups and the remote groups.

We also made some minor changes to the method. In particular, final test performance in Experiment 1 was only assessed by four problems; in Experiment 2, we increased the number of problems to eight in an attempt to obtain more stable estimates of test performance. We also counterbalanced the type of problem that participants practiced first and later

Fig. 4 Mean final test performance for wedge problems in Experiment 1. Error bars represent the standard error of the mean

attempted to solve first during the test. Half the participants practiced and solved wedge problems first, whereas the other half practiced and solved spheroid problems first. Wedge and spheroid problems were selected for counterbalancing because they were the normatively easy and difficult problem types (respectively), and hence we could investigate whether problem difficulty influences the impact of the practice schedules. Indeed, the effectiveness of interleaved practice has been shown to vary by task difficulty (e.g., Wulf & Shea, 2002), but it is unclear how it will affect performance in the present context.

Method

Participants, design, and materials

Participants ($N = 214$) were recruited from Kent State University to fulfill a partial course requirement and randomly assigned to one of four practice groups: blocked, interleaved, remote blocked, and remote interleaved. Additionally, approximately half of the participants in each group practiced and were tested on either the easiest volume problem (i.e., wedge) or the most difficult volume problem (i.e., spheroid). The order for the blocked and interleaved group for half of the participants was switched for the volume wedge and volume spheroid problems, and in the remote practice groups participants only practiced and were tested on either wedges or spheroids. A power analysis for a factorial ANOVA, with power set at .08, $\alpha = .05$, and number of groups set to four, indicated that a sample size of $N = 128$ was sufficient to detect medium-sized effects ($f > .25$). We increased the sample size to $N = 214$ to account for the additional groups that resulted from counterbalancing the order of problem difficulty. Thirteen participants did not attend the final test session, resulting in a final $N = 201$: blocked ($n = 53$), interleaved ($n = 50$), remote-blocked ($n = 48$), and remote-interleaved ($n = 50$).

Procedure

The procedure was identical to Experiment 1 except for the counterbalancing of difficulty mentioned above, the inclusion of the remote-blocked practice group, and an increase in the number of novel test items. As noted in Table 1, the remote-blocked group studied volume, fraction addition, exponent division, and permutation problems in a fixed block order. On the final test participants were given eight novel instances of each problem type. Similar to Experiment 1, the blocked order at test corresponded to the order in which participants practiced solving each problem type, which differed based on whether wedges or spheroids were practiced first.

Results and discussion

We present the results of the pretest and the practice phase followed by analysis of formula retrieval and overall performance on the final test. For formula retrieval and final test performance, we first present analyses of the critical problems (i.e., wedges and spheroids) followed by analyses of all problems.

Pretest performance and practice performance

A one-way ANOVA on performance of all problems on the pretest indicated no significant difference between the blocked group ($M = .31$, $SD = .11$), the interleaved group ($M = .30$, $SD = .12$), the remote-interleaved group ($M = .27$, $SD = .14$), and the remote-blocked group ($M = .31$, $SD = .13$), $F(3,197) = .37$, $MSE = .016$, $p = .78$, $\eta^2 = .005$. Similarly, when analysis was restricted to the geometric shape problems on the pretest, no significant differences occurred between the blocked ($M = .00$, $SD = .00$), interleaved ($M = .00$, $SD = .00$), remote-interleaved ($M = .00$, $SD = .00$), and remote-blocked ($M = .01$, $SD = .07$) groups, $F(3,197) = 2.17$, $MSE = .001$, $p = .09$, $\eta^2 = .032$.

We first analyzed practice performance for the critical problems. Practice performance for the critical problems was not significantly different for blocked group ($M = .81$, $SD = .29$) compared to the interleaved group ($M = .72$, $SD = .35$), $t(101) = 1.53$, $p = .07$, $d = .30$. Practice performance for the critical problems was not significantly different for the remote-blocked group ($M = .89$, $SD = .33$) compared to the remote-interleaved group ($M = .84$, $SD = .25$), $t(96) = .12$, $p = .45$, $d = .02$. As in Experiment 1, we also evaluated practice performance for all volume problems for the interleaved ($M = .68$, $SD = .32$) and blocked groups ($M = .90$, $SD = .13$), $t(101) = 4.59$, $p < .001$, $d = .90$, which demonstrated a significant decline in practice performance for the interleaved group.

Formula retrieval

The mean proportion of correct formula retrieval for critical problems is plotted in Fig. 5. We conducted a factorial ANOVA on formula retrieval for critical problems (i.e., wedges and spheroids), using schedule (blocked vs. interleaved) and remoteness (standard vs. remote) as between-subjects factors. The main effect of schedule was significant, $F(1,197) = 8.33$, $MSE = .206$, $p = .004$, $\eta^2_p = .04$, indicating that formula retrieval was significantly greater for interleaved than blocked groups. The main effect of remoteness was also significant, $F(1,197) = 6.54$, $MSE = .206$, $p = .011$, $\eta^2_p = .03$, indicating that formula retrieval was significantly greater for remote than standard groups. Critically, the schedule by remoteness interaction was not significant, $F(1,197) = .003$, $MSE = .206$, $p = .954$, $\eta^2_p < .001$, indicating that the effect of interleaving on critical problem formula retrieval was not

Fig. 5 Mean proportion of correct formula retrieval for critical (i.e., wedge and spheroid) problems during final test in Experiment 2. Error bars represent the standard error of the mean

different for the standard and remote groups. To test for evidence for the null effect of the schedule by remoteness interaction, we computed a Jeffreys-Zellner-Siow (JZS) Bayes factor using the ANOVA function in JASP (Love et al., 2015; Morey & Rouder, 2015; Rouder, Morey, Speckman, & Province, 2012). A JZS Bayes factor ANOVA with default prior scales indicated a preference for main effects models to the interaction model by a Bayes factor of 4.60. Thus, the data provided *substantial* evidence in favor of a null effect of the interaction between schedule and remoteness on formula retrieval for critical problems (Jarosz & Wiley, 2014).²

Accuracy of formula retrieval for all geometric shapes is presented in the third row of Table 2. A factorial ANOVA on formula retrieval for all geometric shapes revealed a significant main effect of schedule, $F(1,197) = 8.98$, $MSE = .154$, $p = .003$, $\eta^2_p = .04$, and remoteness, $F(1,197) = 19.66$, $MSE = .154$, $p < .001$, $\eta^2_p = .09$. The interaction was not significant, $F(1,197) = .12$, $MSE = .17$, $p = .68$, $\eta^2_p < .001$.

Final test performance

Next, we tested whether the practice schedule for each group affected final test performance, beginning with final test performance for the critical problems (i.e., wedge and spheroid). The mean final test performance for critical problems for each group are displayed in Fig. 6. We conducted a factorial ANOVA on final test performance for critical problems, using schedule and remoteness as between-subjects factors. The main effect of schedule was not significant, $F(1,197) = 3.72$, $MSE = .165$, $p = .06$, $\eta^2_p = .02$. The main effect of remoteness was also not significant, $F(1,197) = 3.41$, $MSE = .165$, $p = .07$, $\eta^2_p = .02$.

² As described in Jarosz and Wiley (2014), Bayes factors between 3 and 10 are thought to provide *positive* (cf. Raftery, 1995) and *substantial* (cf. Jeffreys, 1961) evidence in support for the null hypothesis.

Critically, the schedule by remoteness interaction was not significant, $F(1,197) = .17$, $MSE = .165$, $p = .68$, $\eta^2_p < .001$. Given that this lack of an interaction is relevant to our most critical competitive evaluation of the two hypotheses, we also tested for evidence for the null effect of the schedule by remoteness interaction by computing a JZS Bayes factor. A JZS Bayes factor ANOVA with default prior scales indicated a preference for main effects models to the interaction model by a Bayes factor of 4.45. Thus, the data provided *substantial* evidence in favor of a null effect of the interaction between schedule and remoteness on final test performance for critical problems (Jarosz & Wiley, 2014).

The mean final test performance for all problems for each group is displayed in the fourth row of Table 2. A factorial ANOVA on final test performance for all problems revealed a significant main effect of schedule, $F(1,197) = 5.19$, $MSE = .058$, $p = .02$, $\eta^2_p = .03$, and a significant main effect of remoteness, $F(1,197) = 154.00$, $MSE = .058$, $p < .001$, $\eta^2_p = .44$. The interaction was not significant, $F(1,197) = 1.72$, $MSE = .058$, $p = .19$, $\eta^2_p = .008$.

In summary, interleaving practice – compared to blocking practice – resulted in better performance on all outcome measures except for the final test performance of the critical problems. The advantages of interleaving over blocking were observed in both standard and remote practice schedules. Support for this conclusion comes from evidence of a null effect of the interaction between schedule and remoteness in all analyses.

Analysis of easy versus difficult critical problems

Given that we had chosen a normatively difficult and easy problem as materials in Experiment 2, we examined formula retrieval and final test performance (a) for the participants who studied the normatively easy problems in the first block (i.e., wedges) and (b) for the participants who studied the

Fig. 6 Mean final test performance for critical problems (wedge and spheroid) in Experiment 2. Error bars represent the standard error of the mean

normatively difficult problems first (i.e., spheroids). We conducted a factorial ANOVA on formula retrieval for critical problems, using difficulty, schedule, and remoteness as between-subjects factors. The means are displayed in Fig. 7. The main effect of schedule was significant, $F(1,193) = 8.11$, $MSE = .20$, $p = .005$, $\eta^2_p = .04$, and the main effect of remoteness was also significant, $F(1,193) = 7.36$, $MSE = .20$, $p = .007$, $\eta^2_p = .04$. The main effect of difficulty was not significant, $F(1,193) = 3.34$, $MSE = .20$, $p = .07$, $\eta^2_p = .02$. The remoteness by difficulty interaction was significant, $F(1,193) = 4.47$, $MSE = .20$, $p = .04$, $\eta^2_p = .02$. However, the remaining two-way interactions (schedule by remoteness, schedule by difficulty), and the three-way interaction (schedule by remoteness by difficulty) were not significant, $F_s < 1$.

To follow up on the remoteness by difficulty interaction, we compared the formula retrieval of remote groups to the

standard groups for each level of difficulty. For the easy problems, formula retrieval was not significantly different between standard groups ($M = .39$, $SD = .48$) and remote groups, ($M = .43$, $SD = .49$), $t(103) = .42$, $p = .67$, $d = .08$. For the difficult problems, formula retrieval was significantly greater for the remote groups ($M = .45$, $SD = .49$) than for the standard groups ($M = .14$, $SD = .34$), $t(94) = 3.70$, $p < .001$, $d = .76$.

We next conducted a factorial ANOVA on final test performance for critical problems, using difficulty, schedule, and remoteness as between-subjects factors. The means are displayed in Fig. 8. The main effect of schedule was not significant, $F(1,193) = 3.53$, $MSE = .16$, $p = .06$, $\eta^2_p = .02$, and the main effect of remoteness was significant, $F(1,193) = 4.33$, $MSE = .16$, $p = .04$, $\eta^2_p = .02$. The main effect of difficulty was also significant, $F(1,193) = 5.69$, $MSE = .16$, $p = .02$, $\eta^2_p = .03$. The remoteness by difficulty interaction was significant, $F(1,193) =$

Fig. 7 Mean proportion of correct formula retrieval during final test for easy (i.e., wedge) versus difficult (i.e., spheroid) problems in Experiment 2. Error bars represent the standard error of the mean

Fig. 8 Mean final test performance for easy (i.e., wedge) versus difficult (i.e., spheroid) problems in Experiment 2. Error bars represent the standard error of the mean

8.07, $MSE = .16$, $p = .005$, $\eta^2_p = .04$. However, the remaining two-way interactions (schedule by remoteness, schedule by difficulty), and the three-way interaction (schedule by remoteness by difficulty) were not significant, $F_s < 1$.

To follow up on the remoteness by difficulty interaction, we compared final test performance of the remote groups to the standard groups for each level of difficulty. For the easy problems, final test performance was not significantly different between standard groups ($M = .40$, $SD = .45$) and remote groups, ($M = .35$, $SD = .43$), $t(103) = .50$, $p = .62$, $d = .10$. For the difficult problems, formula retrieval was significantly greater for the remote groups ($M = .38$, $SD = .40$) than for the standard groups ($M = .10$, $SD = .26$), $t(94) = 4.05$, $p < .001$, $d = .83$.

General discussion

Does interleaving versus blocking practice improve problem solving of math concepts? We observed benefits of interleaving practice on formula retrieval and on final test performance of geometric shapes in Experiments 1 and 2. These results replicate research showing benefits of interleaving for math materials (Rohrer & Taylor, 2007; Taylor & Rohrer, 2010; Rohrer, 2012; Sana et al., 2017). Most important, relevant to explaining such interleaving effects, evidence from the current experiments provides more support for the distributed-practice hypothesis than for the discriminative-contrast hypothesis.

The discriminative-contrast hypothesis states that interleaved practice should produce improved learning on criterion tests because instances of different problem types within the same problem domain are practiced in close proximity, which encourages participants to notice similarities and differences across problem types (Carvalho & Goldstone, 2014, 2019; Kang & Pashler, 2012; Kornell & Bjork, 2008; Kornell et al.,

2010; Zulkipli & Burt, 2013). From this account, interleaved groups would produce better outcomes than remote-interleaved groups because the interleaved groups practiced computing volumes from different categories of geometric shapes consecutively, whereas the remote-interleaved groups did not. According to the distributed-practice hypothesis, interleaving practice of concepts enhances learning because instances from each problem type are distributed during practice rather than massed (see Cepeda et al., 2006, for a review and meta-analysis of the spacing effect). If so, performance will be greater for the interleaved group (compared to the blocked group) and for the remote-interleaved group (compared to the remote-blocked group). Interleaving effects were demonstrated across both experiments, and, most important, the interleaving effect (in terms of formula retrieval) was the same size (see Fig. 5) regardless of whether discriminative contrast could contribute to performance (the standard groups) or not (the remote groups).

When combining formula retrieval and computation together into a single score (i.e., final test performance), the interleaved and remote-interleaved groups numerically outperformed their blocked controls, but these advantages were not significant (see Fig. 6). We offer the following explanation for these outcomes. Because participants must commit each of the four formulas to memory, the remote-interleaved practice schedule, which capitalizes on the benefits of distributed practice, should produce better encoding of shape formulas. However, even if participants remember the formula, it does not guarantee they will compute the correct answer (cf. Taylor & Rohrer, 2010), and distributed practice would not be expected to improve students' computation skills. Put differently, participants may learn which formula to associate with which shape because of distributed practice, but they answer incorrectly because of an error of multiplication or division.

Of course, the results of the current study do not rule out the possible contribution of discriminative contrast as a partial explanation of interleaving effects. However, interleaving effects may benefit from discriminative contrast more so for tasks involving category learning of easily confused perceptual categories (e.g., bird species, artists' painting styles) and less so for tasks involving retrieval of formulas from verbal labels (e.g., wedge vs. spheroid). Moreover, the Sequential Attention Theory (SAT) proposed by Carvalho and Goldstone (2015, 2019) helps explain when discriminative contrast will contribute more (vs. less) to interleaving effects. According to SAT, interleaved and blocked practice schedules highlight different aspects of to-be-learned material. The practice schedule that highlights the most challenging features of the material should produce better learning. When between-category similarity is high, and it is challenging to discover which features are diagnostic of category membership, interleaving will benefit category learning because similar exemplars from different categories are studied successively. Studying exemplars from different categories successively allows participants to more easily notice the diagnostic features that make categories different from each other. However, when within-category similarity is low and it is challenging to identify features that are shared by exemplars for a given category, blocking is expected to benefit category learning because exemplars from the same categories are studied successively, which can help participants learn the feature that defines the category (Carvalho & Goldstone, 2019). In explaining the current results, the extent to which interleaving benefits learning by engaging a discriminative contrast mechanism should depend on how successful participants are at contrasting shapes from trial to trial. Each shape is distinct (see Fig. 1); that is, between-category similarity is low, so discriminating between the shapes may be a rather trivial task that does not boost learning above what is already known. Instead, the challenge to participants lies in forming associations between each shape and its formula (e.g., spheroid - $\frac{4\pi r^2 h\pi}{3}$). This process is most likely to be facilitated by spacing rather than interleaving.

Discriminative contrast as specified by SAT may help explain our results of performance on easy versus difficult critical problems in Experiment 2. More specifically, the benefits of remote interleaving over blocking on difficult but not easy problems may have emerged because the formula for the difficult problems is more similar to some of the other formulas. Interleaving practice of the difficult problems may have helped participants to notice the differences between the formulas for difficult problems and the formulas from the other problems, thus benefitting their memory for them. By contrast, the formula for the easy problems is unique (i.e., the formula for wedge is the only formula with the coefficient on the numerator of 1 and a denominator of 2), which should

make it easier to discriminate and thus show the least benefit from interleaving. The idea proposed here is that some types of problems may pose challenges due to high between-category similarity in the responses (in this case, responding with the correct formula). This prediction – inspired by SAT (e.g., Carvalho & Goldstone, 2019) – can be tested in future research by manipulating the similarity of the to-be-remembered responses to different categories.

Finally, the present data suggest that only one factor – distributed practice – contributes to the interleaving benefits for retrieving formulas (and solving volume problems). And, although discriminative contrast does contribute to interleaving effects when participants are learning to classify highly confusable (i.e., similar) perceptual stimuli, it appears that distributing practice can contribute here as well. In particular, Birnbaum et al. (2013, Experiment 3) had participants learn to classify species of butterflies that were easily confusable. All practice was interleaved, but the spacing of exemplars from a given species was either large for one group or relatively small for another group. If discriminative contrast is solely responsible for interleaving effects in this context, then both groups will perform similarly. However, final classification performance was greater for the group that had larger than smaller spacing. Along with data from prior studies that demonstrated the contribution of discriminative contrast to learning butterflies, the outcomes overall indicate that both discriminative contrast and distributed practice can benefit learning and performance in this domain.

In conclusion, results from the present experiments offer further evidence that interleaving practice can benefit learning of math materials. As important, the evidence also provides insights into the relative contributions of discriminative contrast and distributed practice to the benefits of interleaving. In particular, participants benefited equally from interleaving whether it occurred in the standard interleaving format (as compared to the standard blocked format) or in the remote interleaving format (as compared to the remoted blocked format). This general boost in performance suggests that the current interleaving effects involving geometry problems is largely (if not solely) driven by distributed practice.

References

- Benjamin, A. S., & Tullis, J. (2010). What makes distributed practice effective? *Cognitive Psychology*, 61(3), 228–247.
- 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.
- Carvalho, P. F., & Goldstone, R. L. (2014). Putting category learning in order: Category structure and temporal arrangement affect the benefit of interleaved over blocked study. *Memory & Cognition*, 42, 481–495.

- Carvalho, P. F., & Goldstone, R. L. (2015). What you learn is more than what you see: What can sequencing effects tell us about inductive category learning? *Frontiers in Psychology*, 6, Article 505.
- Carvalho, P. F., & Goldstone, R. L. (2019). When does interleaving practice improve learning? In J. Dunlosky & K. A. Rawson (Eds.), *Cambridge Handbook of Cognition and Education* (pp. 411–436). New York: Cambridge University Press.
- 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.
- Donovan, J. J., & Radosevich, D. J. (1999). A meta-analytic review of the distribution of practice effect: Now you see it, now you don't. *Journal of Applied Psychology*, 83, 308–315.
- 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, 4–58.
- Faul, F., Erdfelder, E., Lang, A. G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research and Methods*, 39, 175–191.
- Hintzman, D. (1974). Theoretical implications of the spacing effect. In R. L. Solso (Ed.), *Theories in cognitive psychology: The Loyola Symposium* (pp. 77–99). Potomac, MD: Lawrence Erlbaum.
- Hintzman, D. L., Summers, J. J., & Block, R. A. (1975). Spacing judgments as an index of study-phase retrieval. *Journal of Experimental Psychology: Human Learning and Memory*, 104, 31–40.
- Jacoby, L. L. (1978). On interpreting the effects of repetition: Solving a problem versus remembering a solution. *Journal of Verbal Learning and Verbal Behavior*, 17, 649–667.
- Janiszewski, C., Noel, H., & Sawyer, A. G. (2003). A meta-analysis of the spacing effect in verbal learning: Implications for research on advertising repetition and consumer memory. *Journal of Consumer Research*, 30, 138–149.
- Jarosz, A. F., & Wiley, J. (2014). What are the odds? A practical guide to computing and reporting Bayes factors. *Journal of Problem Solving*, 7, 2–9.
- Jeffreys, H. (1961). *Theory of probability* (3rd Ed.). Oxford, UK: 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, 97–103. <https://doi.org/10.1002/acp.1801>.
- Kornell, N., & Bjork, R. A. (2008). Learning concepts and categories is spacing the “Enemy of Induction”? *Psychological Science*, 19, 585–592.
- Kornell, N., Castel, A. D., Eich, T., & Bjork, R. A. (2010). Spacing as the friend of both memory and induction in young and older adults. *Psychology and Aging*, 25, 498–503.
- Love, J., Selker, R., Verhagen, J., Marsman, M., Gronau, Q. F., Jamil, T., Smira, M., Epskamp, S., Wild, A., Morey, R., Rouder, J. & Wagenmakers, E. J. (2015). *JASP* (Version 0.6)[Computer software].
- Mitchell, C., Nash, S., & Hall, G. (2008). The intermixed–blocked effect in human perceptual learning is not the consequence of trial spacing. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 34, 237–242.
- Morey, R. D. & Rouder, J. N. (2015). BayesFactor (Version 0.9.10-2)[Computer software].
- Rafferty, A. E. (1995). Bayesian model selection in social research. *Sociological Methodology*, 25, 111–163.
- Rohrer, D. (2012). Interleaving helps students distinguish among similar concepts. *Educational Psychological Review*, 24(3), 355–367. <https://doi.org/10.1007/s10648-012-9201-3>
- 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, 1323–1330.
- Rohrer, D., Dedrick, R.F., & Stershic, S. (2015). Interleaved practice improves mathematics learning. *Journal of Educational Psychology*, 21, 1323–1330.
- Rohrer, D., & Taylor, K. (2007). The shuffling of mathematics practice problems boosts learning. *Instructional Science*, 35, 481–498.
- Rouder, J. N., Morey, R. D., Speckman, P. L., Province, J. M. (2012). Default Bayes factors for ANOVA designs. *Journal of Mathematical Psychology*, 56, 356–374.
- 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.
- Schutte, G. M., Duhon, G. J., Solomon, B. G., Poncy, B. C., Moore, K., & Story, B. (2015). A comparative analysis of massed vs. distributed practice on basic math fact fluency growth rates. *Journal of School Psychology*, 53, 149–159.
- Taylor, K., & Rohrer, D. (2010). The effects of interleaved practice. *Applied Cognitive Psychology*, 24, 837–848.
- 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, 750–763. <https://doi.org/10.3758/s13421-010-0063-y>.
- Watkins, O. C., & Watkins, M. J. (1975). Buildup of proactive inhibition as a cue-overload effect. *Journal of Experimental Psychology: Human Learning and Memory*, 104, 442–452.
- 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.
- Zulkipli, N. & Burt, J. S. (2013). The exemplar interleaving effect in inductive learning: Moderation by the difficulty of category discriminations. *Memory & Cognition*, 41, 16–27.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.