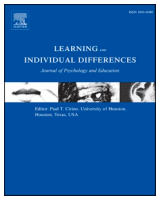
Learning and Individual Differences 117 (2025) 102603

Contents lists available at ScienceDirect

Learning and Individual Differences

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



Individual differences in fluid intelligence moderate the interleaving effect for perceptual category learning☆

Steven C. Pan [a,*], Liwen Yu [a], Yilin Hong [a], Marcus J. Wong [a], Ganeash Selvarajan [a], Michelle E. Kaku [b]

[a] Department of Psychology, National University of Singapore, 9 Arts Link, Singapore City 117572, Singapore
[b] Tokyo Metropolitan Board of Education, JET Programme, 2-8-1 Nishishinjuku, Shinjuku, Tokyo 163-8001, Japan

ARTICLE INFO

Keywords:
Interleaving
Interleaved practice
Fluid intelligence
Episodic memory ability
Working memory capacity
Individual differences

ABSTRACT

The *interleaving effect* refers to the finding that repeatedly switching between categories during study or practice enhances learning relative to focusing on only one category at a time. Two studies investigated whether this effect is moderated by individual differences in fluid intelligence (gF), episodic memory (EM) ability, and/or working memory capacity (WMC). In Study 1 (undergraduate students) and Study 2 (adult online participants), higher gF scores were associated with larger interleaving effects for perceptual categories (artists' painting styles). Additionally, higher EM ability was associated with larger interleaving effects for perceptual categories in Study 1, whereas an analogous pattern was observed for WMC in Study 2. There were no indications that the investigated cognitive abilities moderated the interleaving effect for text-based categories (psychological disorders). Overall, these findings suggest that higher-ability learners benefit especially from interleaving in the case of perceptual category learning, with attendant theoretical and pedagogical implications.

Educational relevance and implications statement

A growing body of research suggests that repeatedly switching between to-be-learned categories as they are learned, or *interleaving*, can benefit learning in a variety of circumstances. The extent to which that phenomenon, which is formally known as the *interleaving effect*, may vary among learners of differing cognitive abilities has yet to be fully explored. We found that the interleaving effect for learning perceptual categories (artists' painting styles) varied based on fluid intelligence (the capacity to reason, think abstractly, and solve problems), with individuals scoring higher on measures of fluid intelligence exhibiting a larger interleaving effect. Moreover, there were also some indications that individuals scoring higher on measures of episodic memory (the ability to remember prior experiences) or working memory capacity (the ability to focus on memories that are relevant to a goal or task at hand) also exhibited a larger interleaving effect. Together, these results suggest that interleaving may be an especially effective learning tool for individuals with higher cognitive ability scores, while still offering benefits to those with lower scores.

1. Individual differences in fluid intelligence moderate the interleaving effect for perceptual category learning

In many educational contexts, students engage in inductive learning (i.e., learning from examples) of a series of categories. For instance, in geology classes, different types of rocks are often learned, whereas in many clinical courses, various health conditions are learned. What is the optimal way to arrange learning activities in such situations? A popular and intuitive approach is to engage in *blocking* (or *blocked practice*), which involves focusing on one category at a time (e.g., studying multiple examples of igneous rocks before moving on to metamorphic rocks). A lesser-known alternative, *interleaving* (or *interleaved practice*),

☆ Thanks to Aruna Kandasamy, Josiah Hoo, Kaigene Chen, Nathaneal Teo, and Zihan Cui for assistance with data collection, Andy Teo for help with experimental programming, and Jolynn Pek for analysis suggestions. Thanks also to Nate Kornell, Gene Brewer, and Alexander Burgoyne for sharing stimulus materials, as well as Faria Sana for assistance with software resources. This research was supported by a National University of Singapore Faculty of Arts & Social Sciences (FASS) grant to S. C. Pan.Data and analysis code are available at the Open Science Framework (OSF) and can be accessed at https://osf.io/ng8wd/, whereas materials are accessible at https://osf.io/dqprv/

* Corresponding author at: Department of Psychology, Faculty of Arts and Social Sciences, National University of Singapore, 9 Arts Link, Singapore City 117572, Singapore.
E-mail address: scp@nus.edu.sg (S.C. Pan).

https://doi.org/10.1016/j.lindif.2024.102603
Received 28 May 2024; Received in revised form 27 October 2024; Accepted 26 November 2024

involves switching between categories repeatedly during learning (e.g., studying randomly intermixed examples of different rock types). Interleaving is often more difficult and challenging for students to use, at least initially, but can ultimately yield better learning than blocking. That phenomenon is known as the *interleaving effect* (for reviews see Carpenter, 2014; Carpenter & Pan, 2024; Carvalho & Goldstone, 2017; Kang, 2017; Rohrer, 2012); for meta-analyses, see Brunmair & Richter, 2019; Firth et al., 2021).

Studies of the interleaving effect usually involve the study of category exemplars using interleaving or blocking (e.g., examples of landscape artists' painting styles as in Kornell & Bjork, 2008; images of bird families as in Wahlheim et al., 2011). With interleaving, there is repeated alternation between categories such that each study or practice attempt addresses a different category than the previous one. In contrast, blocking concentrates all study or practice attempts for each category together, moving on to the next category only after the current one is completed. After a delay ranging from a few seconds to several days, learning is assessed via a classification test or other assessment, and on that test, an interleaving effect—i.e., better performance for interleaved versus blocked materials—is usually observed for most participants. According to Brunmair and Richter (2019), the interleaving effect varies in magnitude depending on the types of materials (i.e., categories) being learned; it is most potent for learning artists' painting styles (Hedges' $g = 0.67$) and is less potent for naturalistic photographs ($g = 0.36$) or expository texts ($g = 0.21$). The interleaving effect also tends to be stronger when the categories being learned are relatively similar or confusable with one another.

1.1. Theoretical accounts of the interleaving effect

Multiple theoretical accounts have been proposed to explain the interleaving effect (for reviews see Brunmair & Richter, 2019; Carvalho & Goldstone, 2019). Arguably the most prominent explanation is the *discriminative contrast hypothesis*, which posits that interleaving enables learners to compare and contrast different categories, thereby enhancing their ability to recognize and differentiate categories. First proposed by Kang and Pashler (2012) and later refined by Carvalho and Goldstone (2017) as the *sequential attention theory* (which suggests that interleaving and blocking focus learners' attention on between-category and within-category features, respectively), it is consistent with the findings that the interleaving effect is strongest for very similar or confusable categories (Brunmair & Richter, 2019), interrupting interleaving with filler tasks reduces learning outcomes versus conventional interleaving (e.g., Birnbaum et al., 2013; Zulkiply & Burt, 2013), and that the simultaneous presentation of different categories—presumably facilitating comparison—yields learning outcomes similar to that of interleaving (e.g., Kang & Pashler, 2012; see also Sana et al., 2017).

Another class of accounts attributes the interleaving effect to the well-established *spacing effect* (Ebbinghaus, 1885; Kornell & Bjork, 2008; see also Carpenter et al., 2022; Carpenter & Pan, 2024; Cepeda et al., 2006; Delaney et al., 2010), wherein learning opportunities that are spread out (or "spaced") in time enhances memory. Spacing is an inherent property of interleaving as each category is not revisited until other categories are introduced first. Spacing effect-based explanations for the interleaving effect include the *attention attenuation hypothesis* (e.g., Wahlheim et al., 2011), wherein learner attention is reduced during blocking compared to interleaving (see also the *deficient processing* account; e.g., Toppino & Bloom, 2002), as well as the *study-phase retrieval* account (e.g., Hintzman et al., 1975), wherein interleaving prompts retrieval of information about the last exposure to a to-be-learned category from long-term memory, yielding improved learning.

The foregoing perspectives are not exhaustive and theoretical development pertaining to the interleaving effect remains a work in progress. To date, no hypothesis or account of the effect has received universal support. On one hand, studies with filler tasks suggest that the interleaving effect relies more on discriminative contrast than any processes stemming from a spacing effect (e.g., Birnbaum et al., 2013; Kang & Pashler, 2012; Taylor & Rohrer, 2010); conversely, interleaving effects have been observed in cases where the to-be-learned materials are not highly similar or confusable (e.g., Rohrer et al., 2014; see also Foster et al., 2019), challenging a purely discriminative contrast explanation. Another possibility is that the interleaving effect relies on multiple mechanisms that play greater or lesser roles depending on the materials being learned.

1.2. Individual differences and the interleaving effect

The growing evidence showing the interleaving effect in such disciplines as chemistry education (e.g., Eglington & Kang, 2017), introductory physics (e.g., Samani & Pan, 2021), second language learning (e.g., Suzuki et al., 2022), source evaluation (e.g., Abel, Roelle, & Stadtler, 2024), and other areas has led some learning scientists to endorse interleaving as a "desirable difficulty" that students should use regularly (e.g., Bjork & Bjork, 2011; see also Firth et al., 2021). The extent to which interleaving benefits different learners, however, remains unclear (for related discussion, see Abel et al., 2021). Students vary in established cognitive abilities such as *fluid intelligence* (gF; the ability to think abstractly and solve problems; Kievit et al., 2016), *episodic memory ability* (EM; the extent to which one can remember specific items and contextual information from prior experiences; Blankenship et al., 2015), and *working memory capacity* (WMC; the ability to focus on goal-relevant memories, particularly in the face of distraction; Engle & Kane, 2003), all of which can explain patterns in academic performance (Alloway & Alloway, 2010; Blankenship et al., 2015; Colom et al., 2007; Di Fabio & Busoni, 2007; Ren et al., 2015; see also Unsworth, 2016). Higher gF, EM ability, and/or WMC scores tend to be associated with better learning outcomes (Kyllonen & Christal, 1990; Shipstead et al., 2016).

Various theories have explored the relationships among gF, EM ability, and/or WMC (e.g., Engle & Kane, 2003). Strong correlations between gF and WMC have been observed (e.g., Kyllonen & Christal, 1990), with some studies suggesting that this relationship is mediated by long-term memory processes (i.e., EM ability) and attention control (e.g., Unsworth & Spillers, 2010). Hence, regarding these cognitive abilities as entirely independent would overlook their potential interconnections. A common approach in the field is to treat gF, EM ability, and WMC as distinct yet interrelated constructs (e.g., Brewer & Unsworth, 2012; see also Robey, 2019). This approach is buttressed by studies suggesting that these abilities make nonredundant contributions to cognition. For example, tasks involving EM ability versus gF show nonidentical neural activation patterns (Raykov et al., 2024), and factors such as aging and targeted interventions (e.g., physical exercise) affect these abilities in distinct ways (Kachouri et al., 2022; Salthouse et al., 2008).

Due to its presumed role in managing and comparing features across categories, supporting long-term memory retrieval, and resisting distraction (see Wang et al., 2020)—the latter possibly relating to attention-based accounts of interleaving effects—most research on individual differences in interleaving to date has focused on WMC, typically measured via operation span tasks (Unsworth et al., 2005; cf. Suzuki et al., 2022). Sana et al. (2017) found that lower-WMC individuals benefited more from interleaving when learning nonparametric statistical procedures from short texts, performing as well as higher-WMC individuals under interleaving but worse under blocking; however, this finding was not fully replicated in Sana et al. (2018; Experiment 1). Studies with perceptual categories have also found no moderating effect of WMC, even under dual task conditions (e.g., Sana et al., 2018, Experiments 2–5; Wang et al., 2020; Yan & Sana, 2021), although Guzman-Munoz (2017; Experiments 2 and 3) reported marginal evidence for a larger interleaving effect among higher-WMC individuals.

Unlike WMC, the influence of gF and EM ability on the interleaving

effect has received little investigation, but there are reasons to suspect that they may play a moderating role. For instance, individuals with higher gF and EM scores tend to use more effective memory strategies such as imagery, the keyword method, and elaboration (e.g., Kirchhoff, 2009; Minear et al., 2018; see also Bailey, Dunlosky, & Kane, 2008; Robey, 2019), which could influence how much they benefit from interleaving. Moreover, if these strategies enhance retention and retrieval of information across interleaved categories—a process implicated in spacing effect-based accounts—then the interleaving effect may be affected. The advanced abstract reasoning abilities of higher-gF individuals (Carroll, 1993) might improve their ability to extract salient features during interleaving, potentially affecting effective discriminative contrast processes. In a similar vein, EM ability could impact the ability to remember distinguishing features, further influencing the interleaving effect. All of these possibilities, however, have yet to be explored.

To our knowledge, no studies to date have addressed EM ability and the interleaving effect. Only a study by Del Missier et al. (2018; Experiment 2) involving word list learning has addressed gF (via Raven's standard progressive matrices) and the interleaving effect. They found that gF scores were positively correlated with test performance in a blocked condition but not in an interleaved condition. Given the study's focus on generating proactive interference through the study of multiple word lists, however, it is unclear whether its findings generalize to the wider literature on the interleaving effect.

1.3. The current study

We investigated individual differences in gF and EM ability and their relationship with the interleaving effect. Our investigation comprised two studies of nearly identical design, sampling from different populations in different geographical locations: undergraduate students at a large public research university (Study 1) and adults of various ages and educational backgrounds sampled online (Study 2). Sampling from two different populations not only enabled us to reach a potentially broader range of abilities than a single study alone, but also allowed us to address the potential reproducibility of any observed moderating effects of cognitive abilities on the interleaving effect.

All participants completed two interleaving learning tasks, one based on Kornell and Bjork's (2008) widely-cited study of perceptual category learning (with artists' painting styles) and the other based on Zulkiply et al.'s (2012) study of text-based category learning (involving psychological disorders presented in case study format). Including two tasks allowed us to address the interleaving effect in two different domains and with two modalities where the typical effect size of the interleaving effect, and possibly associated cognitive processes, differ (Brunmair & Richter, 2019). Both tasks have educational relevance: Perceptual category learning commonly occurs in such contexts as physical sciences courses, whereas learning about categories from text-based materials frequently occurs throughout many levels of education. Examples of the stimulus materials used in both tasks are presented in Fig. 1.

In alignment with prior studies of individual differences in cognitive abilities and learning strategies (e.g., Brewer & Unsworth, 2012; Pan et al., 2015; Robey, 2019), all participants completed three tasks each to measure gF and EM ability. Using several tasks to measure a cognitive ability is a recommended approach in individual differences research (for discussions see Engle & Kane, 2003; Unsworth, 2019; Wingert & Brewer, 2018). As a secondary measure of interest, however, each participant also completed a WMC task (which enabled comparisons with prior research on WMC and the interleaving effect).

For each investigated cognitive ability, we considered three possible outcomes. These outcomes can be visualized as hypothetical plots of cognitive ability scores versus classification test scores on an interleaving effect task, plotted separately for interleaved and blocked conditions. Those plots are featured in Fig. 2, in which potential moderating effects of a cognitive ability on the interleaving effect can be defined as scenarios in which that ability influences performance in the interleaved condition, the blocked condition, or both. The three outcomes can be characterized as follows:

1. *Higher-ability learners benefit most from interleaving.* A manifestation of this outcome entails a steeper positive slope in the interleaved condition than in the blocked condition. If so, then it would suggest that higher-ability learners derive greater benefits from interleaving than lower-ability learners, whereas blocking yields more similar levels of learning across the ability range. The net result is a larger magnitude interleaving effect for higher-ability learners. This outcome could be described as interleaving helping the "rich get richer."
2. *Lower-ability learners benefit most from interleaving.* This outcome could manifest as a steeper positive slope in the blocked condition than in the interleaved condition. In this case, lower-ability learners exhibit poorer learning than higher-ability learners when blocking is

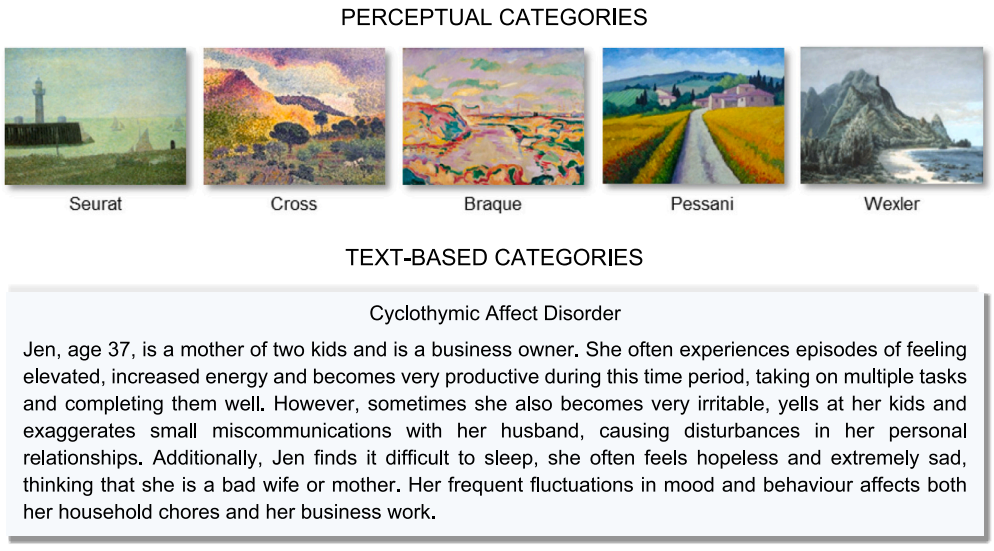


Fig. 1. Example perceptual and text-based interleaving task stimuli.
Note: The perceptual categories encompassed a total of 12 artists and the text-based categories encompassed a total of 6 different psychological disorders. Cyclothymic disorder, an example of a text-based category that is featured in the figure, is typically classified as a milder form of bipolar disorder.

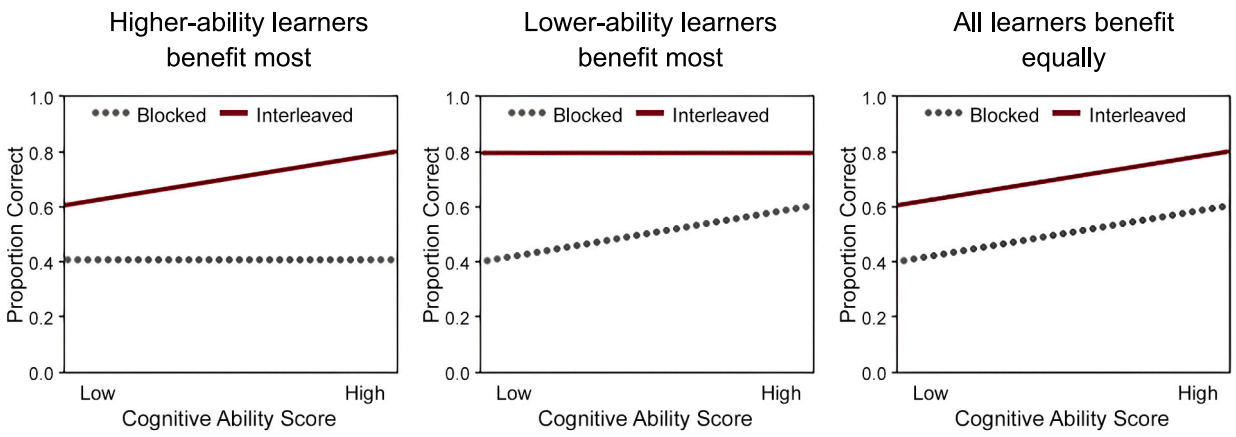


Fig. 2. Hypothetical relations between cognitive abilities and the interleaving effect.
Note: The above panels display hypothetical performance on a classification task (after materials have been learned through blocking and interleaving) as a function of cognitive ability scores. Depending on the scenario, performance in the blocked and interleaved conditions may or may not differ among individuals of low versus high cognitive ability (e.g., gF, EM ability).

used, whereas all learners perform more similarly when interleaving is used. Consequently, a larger interleaving effect is observed for lower-ability learners. With this outcome, interleaving "levels the playing field" for learners.

3. *All learners benefit from interleaving equally*. This outcome would be observed if the slopes in the blocked and interleaved conditions do not substantially differ. In such a scenario, interleaving effect magnitude is similar across the ability range. Under this outcome, the interleaving effect acts as "a rising tide lifts all boats."

It should be noted that the literature on individual differences and interleaving to date, which has focused on WMC, has largely reported findings that are consistent with the third outcome. Whether the same patterns would be observed for the case of gF and EM ability, however, was unknown prior to this investigation.

2. Study 1

The initial study investigated whether individual differences in cognitive abilities—specifically gF, EM ability, and WMC—moderate the benefits of interleaved practice for perceptual and/or text-based category learning among a sample of undergraduate students.

2.1. Methods

2.1.1. Participants

The minimum sample size for both studies, 120, was comparable to sample sizes used in previous research on individual differences in WMC and the interleaving effect (e.g., Sana et al., 2017, 2018; Wang et al., 2020). Data collection occurred over a five-month academic semester at a large public research university in Singapore. We aimed to exceed the minimum sample size target, and by semester's end, 167 undergraduate students from the psychology subject pool at the university had participated in exchange for partial course credit. All participants were either native English speakers or highly fluent in English and had not previously studied or had extensive experience with artists' painting styles or psychological disorders. Data from five participants were excluded due to experiencing technical problems and/or noncompliance with study instructions. The final sample of 162 participants had a mean age of 21.5 years (range: 18 to 42 years) and was 63.6 % female. Reflecting the location in which the study occurred, the vast majority of participants were Asian, with ethnic backgrounds comprised of 79.2 % Chinese, 11.0 % Indian, 4.5 % Malay, and 5.2 % of other ethnicities. The sample was dominated by first-year students (61.1 %; with 16.0 % and 22.8 % of the students being in their second or at least third year of study, respectively). The three most represented academic majors were Psychology (21.6 %), Nursing (13.0 %), and Business (10.5 %).

Data collection for all studies reported in this manuscript was conducted with ethics approval obtained from the same university on July 11, 2023 (Protocol 2023-June-09, amended 2023-Nov-17). All participants provided informed consent beforehand. Each participant was treated in compliance with the principles outlined in the Declaration of Helsinki.

2.1.2. Materials

2.1.2.1. Interleaving learning tasks. Materials for the two interleaving tasks were as follows.

2.1.2.1.1. Perceptual categories. Materials were drawn verbatim from Kornell and Bjork (2008) and consisted of ten example paintings each from the artists Georges Braque, Henri-Edmond Cross, Judy Hawkins, Philip Juras, Ryan Lewis, Marilyn Mylrea, Bruno Pessani, Ron Schlorff, Georges Seurat, Ciprian Stratulat, George Wexler, and YieMei. Six examples per artist were used during the initial study phase (72 images total), whereas four examples per artist were used during the subsequent classification test (48 images total). Half of the artists were learned using interleaving, whereas the other half were learned using blocking; the assignment of artists to interleaving or blocking was identical for all participants. Example stimuli are presented in Fig. 1.

2.1.2.1.2. Text-based categories. Materials were six case studies each describing attention deficit hyperactivity disorder, autism spectrum disorder, bipolar disorder, borderline personality disorder, intellectual development disorder, and schizophrenia. The case studies were based on examples included in Zulkiply et al. (2012) (see also Murphy & Pavlik, 2018). All case studies consisted of a single paragraph of 100–120 words in length describing an individual with behavioral characteristics that met the diagnostic criteria for the respective disorder. Three case studies per disorder were presented during the initial study phase (18 case studies total), whereas the remaining three case studies per disorder were used during the subsequent classification test (9 case studies total). To help mitigate potential order effects (and considering the relatively small number of categories and case studies), the assignment of categories to be interleaved or blocked during initial study occurred using three counterbalancing schemes, with each participant receiving one of those schemes. An example case study is presented in Fig. 1.

As the original case study materials were no longer available (N. Zulkiply, personal communication, July 26, 2019), newly constructed versions were used that resembled those materials as closely as possible but with one major change to better align with authentic learning

contexts: Whereas Zulkiply et al. (2012) used nonsense names (e.g., "Hix") in place of the actual disorder name, we used an obscure but clinically relevant name (i.e., cyclothymic affect disorder, dysfunctional cognition disorder, pervasive development disorder, resonance development disorder, schismic cognition disorder, and self-regulation disorder) to increase the plausibility of the materials.

2.1.2.2. Fluid intelligence tasks. The tasks assessing gF (Raven's progressive matrices, number series, and letter sets tasks) used materials previously developed for these assessments. Raven's progressive matrices included 18 test items (Raven & Raven, 2003), the number series task consisted of 15 test items (Ekstrom et al., 1976), and the letter sets task comprised 20 test items (Thurstone, 1938).

2.1.2.3. Episodic memory tasks. The tasks assessing EM ability (delayed free recall, cued recall, and recognition tasks) used materials from Brewer and Unsworth (2012). The delayed free recall task featured six lists of 10 words each, the cued recall task involved three lists of 10 word pairs each, and the recognition task consisted of 60 drawings depicting diverse objects.

2.1.2.4. Working memory task. The WMC task was a version of the operation span (Unsworth et al., 2005) that is commercially available via the Inquisit (Millisecond Software, 2023) online software platform. The constituent materials in that task, namely letters and math problems, were used without modification.

2.1.3. Procedure and scoring

All participants underwent a single session lasting approximately 90 min in a laboratory testing room. They used desktop PCs or docked laptop computers equipped with the Google Chrome internet browser and Inquisit Web 5 software while situated at individual laboratory testing cubicles or desks. After providing informed consent and responding to demographic questions, they completed the two interleaving tasks in counterbalanced order (with either the perceptual or the text-based task first). Following these initial tasks, they completed the operation span task, the three gF tasks (Raven's progressive matrices, number series, and letter sets), and the three EM tasks (delayed free recall, cued recall, and image recognition), in that order. Participants were allowed brief breaks between tasks if desired. After completing the final task, they received a debriefing and were dismissed.

2.1.3.1. Interleaving learning tasks. We used the visual category learning task popularized by Kornell and Bjork (2008) and a text-based category learning task developed by Zulkiply et al. (2012). As all participants indicated no prior experience with artists' painting styles or psychological disorders—and were encountering these concepts for the first time through the examples presented in the tasks, without any additional explanations or training—they were engaging entirely in inductive learning as they performed the tasks. Details of the tasks are as follows.

2.1.3.1.1. Perceptual categories. During an initial study phase, participants viewed six examples from each of 12 different artists (for a total of 72 examples). Each painting was displayed one at a time for 3 s with the last name of the artist displayed below. Each consecutive block of six paintings comprised either six paintings by a single artist (blocked) or one painting by each of six artists (interleaved). As in Kornell and Bjork (2008), the order of the blocks followed the pattern BIIBBIIBBIIB (where B = blocked and I = interleaved). The order of paintings within each six-painting block was randomized anew for each participant. Next, participants paused for 30 s before proceeding to the classification test. On the test, four new examples of each artist were presented in random order. Each example was accompanied by a list of the 12 artists, with participants required to select the artist that they believed painted that example. Feedback was not provided. The test was self-paced and ended once participants had finished answering all 48 items. The interleaving effect was computed for each participant by subtracting the mean classification test score for all blocked artists from the mean test score for all interleaved artists.

2.1.3.1.2. Text-based categories. During the initial study phase, participants viewed three example case studies for each of six different disorders (for a total of 18 examples). Each case study was presented one at a time for 30 s in text format with the name of the disorder displayed above. Each consecutive block of three case studies comprised either three examples of the same disorder (blocked) or one example of each of three different disorders (interleaved). As in Zulkiply et al. (2012), the order of the blocks was BIBIBI (where B = blocked and I = interleaved). The order of case studies within each block was randomized anew for each participant. Next, after a break of approximately 30 s, participants proceeded to the classification test. On the test, three new examples of each disorder were presented in random order. Each example was accompanied by a list of the six disorders, with participants required to select the disorder that they believed was described in the example. Feedback was not provided. The test was self-paced and ended once participants had finished answering all 18 items. The interleaving effect was computed for each participant by subtracting the mean classification test score for all blocked disorders from the mean test score for all interleaved disorders.

2.1.3.2. Fluid intelligence tasks. The three fluid intelligence tasks, which included figural matrices and series tests (both widely recognized as effective measures of fluid intelligence; Kyllonen et al., 2017) were as follows.

2.1.3.2.1. Raven's progressive matrices. Participants were allotted 10 min to solve up to 18 logic problems presented in a fixed sequence that progressively increased in difficulty. Each problem featured a 3 × 3 matrix of geometric patterns, with the bottom right pattern missing. From eight options, participants chose the pattern that they believed correctly completed the matrix. Scores were calculated as the proportion of correctly solved problems out of 18.

2.1.3.2.2. Number series. Participants had 5 min to solve up to 15 problems. Each problem presented a sequence of numbers following an undisclosed rule, along with five potential answer options. Problems were presented in the same sequence for all participants. Participants were tasked with selecting the answer option representing the next number in the series. Scores were calculated as the proportion of correctly solved problems out of 15.

2.1.3.2.3. Letter sets. Participants were given up to 5 min to solve up to 20 problems. Each problem featured five sets of four letters, with four sets adhering to an undisclosed rule. Participants were required to identify the set that deviated from this rule. Problems were presented consistently for all participants. Scores were calculated as the proportion of correctly solved problems out of 20.

2.1.3.3. Episodic memory tasks. The three episodic memory tasks, which encompassed free recall, cued recall, and recognition tasks (which are widely used and recommended in psychology research; Cleary, 2018; see also Unsworth, 2019) were as follows.

2.1.3.3.1. Delayed free recall. Participants were presented with six lists, each containing 10 common nouns. Each list went through three phases: list presentation, distractor task, and free recall test. During list presentation, nouns were displayed individually for 1 s each and in the same order for all participants. Next, a 15-s distractor task involved solving three arithmetic problems. Finally, participants underwent a free recall test, where they typed as many words from the most recently presented list as they could remember within a 45-s time frame. Participants' scores were calculated as the proportion of correctly recalled words out of 60, across all six lists.

2.1.3.3.2. Cued recall. Participants learned three lists, each containing 10 word pairs. First, each pair was displayed individually for 2 s

each, in the same order for all participants. Next, a cued recall test required participants to recall and type the missing target word when presented with only the cue word for each pair. This test was self-paced, and no feedback was provided. Participants' scores were calculated as the proportion of word pairs correctly recalled out of 30, across all three lists.

2.1.3.3.3. Image recognition. Participants viewed 30 drawings depicting various common objects, with each image shown individually for 3 s in random order. Subsequently, a recognition test involved viewing 60 drawings individually for 5 s each, also in random order. Participants were tasked with identifying each image as either new or old (including the 30 drawings previously seen and 30 new ones). Their scores were calculated as the proportion of correctly identified images out of 60.

2.1.3.4. Working memory task. Participants completed the operation span task, which measures participants' capacity to retain a sequence of letters in working memory, using Inquisit Web 5. The task, which is a widely used method of measuring WMC (Conway et al., 2005), involved the presentation of randomly ordered sequences containing three to seven letters. Each letter was displayed for approximately 1 s and preceded by a simple math problem. Following the presentation of a given sequence, participants recalled the letters by selecting them in the correct order from a provided letter matrix. The entire task consisted of 15 trials, each featuring a unique sequence of letters. The task was scored using partial credit load scoring, in which each participant's operation span score was the sum of the correctly recalled letters from all sequences, regardless of whether an entire sequence was recalled perfectly (with a maximum possible score of 75). This scoring approach is recommended for its internal consistency (Conway et al., 2005).

2.1.4. Score transformations and analysis plan

R version 4.1.0 (R Core Team, 2021) was used for all data processing and analyses.

2.1.4.1. Composite Z-scores. Following approaches used in prior studies of individual differences (e.g., Brewer & Unsworth, 2012; Robey, 2019; see also Wingert & Brewer, 2018), *z*-score transformations of the gF, EM, and WMC measures were performed (i.e., rescaling the data from each measure to achieve a mean of 0 and a standard deviation of 1). Subsequently, the mean of the respective *z*-scores for each participant was computed to generate composite *z*-scores for gF and EM ability (e.g., the composite EM ability score for a given participant encompassed the average of the *z*-scores derived from the delayed free recall, cued recall, and recognition tasks for that participant). Given that only one task was used to measure WMC, however, a composite score was not computed for WMC.

2.1.4.2. Factor scores. Besides composite *z*-scores, we employed confirmatory factor analysis (CFA) to generate factor scores for gF and EM. Details of this approach are presented in the Supplementary Results.

2.1.4.3. Analysis plan. The analysis plan comprised the following steps. First, we computed descriptive statistics and a correlation matrix for all measures. Second, we investigated the magnitude and variability of the interleaving effect by computing it separately for perceptual categories and text-based categories and identifying the percentages of participants with positive, negative, and null interleaving effects. To further characterize the observed interleaving effects, we also report Bayes factors calculated using the *BayesFactor* package in R (Morey & Rouder, 2012); BF_{10} is reported in cases where the alternative hypothesis is more likely (i.e., $BF_{10} > 3$), and for ease of interpretation, the reciprocal BF_{01} is reported in cases where the null hypothesis is more likely (i.e., $BF_{01} > 1$) (Rouder et al., 2009; Wagenmakers, 2007).

Third, we explored the potential influence of individual differences in cognitive abilities on the interleaving effect by fitting linear mixed-effects models using composite *z*-scores along with classification test data for blocked and interleaved items. Models were fitted separately for perceptual and text-based categories using the lme4 package version 1.1.34 (Bates et al., 2015) in R. Within each set of analyses, we fitted models individually for gF and EM ability to examine interaction effects between learning schedule (blocked vs. interleaved) and the specific individual differences of interest. Each model incorporated learning schedule (blocked = 0 vs. interleaved = 1), the respective composite scores, and their interaction as predictors, with crossed random intercepts for participants. Fourth, for additional insights into the potential role of individual differences on the interleaving effect for perceptual categories, we conducted supplementary analyses that involved dividing the study data into quartiles (cf. Brewer & Unsworth, 2012; Minear et al., 2018). Fifth, to investigate the potential role of WMC on the interleaving effect, we performed linear-mixed effects models using operation span *z*-scores separately for the perceptual and text-based categories (these analyses were less extensive given that WMC was not the main focus of this study).

To provide potentially converging evidence, the third through fourth steps described above were repeated using factor scores (for a similar approach, see Robey, 2019), which offer the advantage of minimizing the measurement error associated with composite scores (Robey, 2019; also see Wingert & Brewer, 2018), for perceptual materials. The results of these analyses, which are reported in the Supplementary Results, were consistent across all comparisons with the composite score-based analyses.

Lastly, we employed structural equation modeling (SEM) to analyze combined data from Studies 1 and 2. These results are presented after the separate results for Study 1 and Study 2.

2.2. Results

Descriptive statistics and split-half reliability values are presented in Table 1 and a correlation matrix of all measures is presented in Table 2. In the following sections, we first present analyses of interleaving effect magnitude and variability, analyses involving gF and EM ability, and then analyses involving WMC.

2.2.1. Interleaving effects

2.2.1.1. Perceptual categories. Participants tended to classify examples

Table 1
Descriptive statistics for Study 1.

Measure	Mean	*SD*	Skewness	Kurtosis	Reliability
Perceptual categories					
Interleaved condition	0.54	0.24	−0.072	−0.98	0.91
Blocked condition	0.30	0.17	0.46	−0.46	0.83
Interleaving effect	0.24	0.19	0.11	−0.25	0.74
Text-based categories					
Interleaved condition	0.62	0.29	−0.40	−1.02	0.82–0.86[a]
Blocked condition	0.61	0.29	−0.33	−1.19	0.79–0.83[a]
Interleaving effect	0.0048	0.24	−0.056	0.67	0.35–0.60[a]
Fluid intelligence (gF)					
Raven's matrices	0.80	0.19	−1.12	0.57	0.86
Letter sets	0.53	0.12	0.12	0.08	0.63
Number series	0.68	0.17	−0.21	−0.43	0.73
Episodic memory (EM) ability					
Delayed free recall	0.50	0.13	−0.10	−0.16	0.80
Cued recall	0.59	0.23	−0.15	−0.95	0.91
Image recognition	0.92	0.091	−1.69	3.08	0.55
Working memory capacity (WMC)					
Operation span	66.03	9.75	−2.82	11.85	0.70

[a] Range of values reflects split-half reliabilities for the different counterbalanced versions of the text-based category learning task.

Table 2
Correlation matrix for Study 1.

Measure	1	2	3	4	5	6	7	8	9	10	11	12	13
Interleaving tasks													
1. I-Visual	1.00												
2. B-Visual	0.63***	1.00											
3. IE-Visual	0.69***	−0.12	1.00										
4. I-Text	0.30***	0.29***	0.12	1.00									
5. B-Text	0.42***	0.28***	0.27***	0.67***	1.00								
6. IE-Text	−0.14	0.0080	−0.19*	0.41***	−0.40***	1.00							
gF tasks													
7. Raven	0.39***	0.23**	0.29***	0.43***	0.43***	0.0023	1.00						
8. Lsets	−0.0080	0.024	−0.032	−0.041	−0.038	−0.0044	0.036	1.00					
9. Nseries	0.25**	0.16*	0.17*	0.28***	0.28***	−0.0042	0.51***	0.21**	1.00				
EM ability tasks													
10. DFR	0.32***	0.32***	0.11	0.25**	0.29***	−0.043	0.32***	0.14	0.29***	1.00			
11. CR	0.43***	0.41***	0.16*	0.30***	0.30***	−0.0070	0.37***	0.0010	0.33***	0.50***	1.00		
12. Recog	0.44***	0.33***	0.25**	0.36***	0.34***	0.035	0.46***	0.10	0.25**	0.34***	0.40***	1.00	
WMC task													
13. Ospan	0.14	0.14	0.048	0.22**	0.22**	−0.0067	0.30***	0.11	0.30***	0.28***	0.21**	0.35***	1.00

Note. * = $p < .05$; ** = $p < .01$; *** = $p < .001$. For the interleaving tasks (1–6): I = interleaved condition; B = blocked condition, IE = interleaving effect; Visual = perceptual categories; Text = text-based categories. For the fluid intelligence tasks (7–9): Raven = Raven's progressive matrices; Lsets = Letter sets; Nseries = Number series. For the episodic memory tasks (10−12): DFR = delayed free recall; CR = cued recall; Recog. = image recognition. Working memory task (13): Ospan = operation span.

from interleaved artists more correctly than examples from blocked artists, $t(161) = 16.37$, $p < .0001$, Cohen's $d = 1.29$, $BF_{10} > 1000$. The mean interleaving effect for perceptual categories was 0.24 proportion correct, which aligns with effects reported in the literature using the same or similar materials. There was variability in this effect: 90.7 % of participants exhibited a numerically positive interleaving effect, 5.6 % exhibited a numerically negative interleaving effect (i.e., classification performance was better for blocked artists), and 3.7 % exhibited no interleaving effect (i.e., equivalent performance for blocked and interleaved artists). Across the entire sample, interleaving effect magnitude ranged from −0.21 to 0.75 proportion correct.

2.2.1.2. Text-based categories. Participants did not classify case studies representing interleaved disorders more correctly than case studies representing blocked disorders, $t(161) = 0.26$, $p = .80$, $d = 0.020$, $BF_{01} = 11.05$. The very small mean interleaving effect for text-based categories, 0.0048 proportion correct, is inconsistent with the results reported by Zulkiply et al. (2012) or another recent study (that used the same case study stimulus materials but in a between-subjects design) in which a significant interleaving effect was observed (Pan, Selvarajan and Murphy, 2024). It does, however, resemble the null results reported by Murphy and Pavlik (2018) using similar materials. Overall, 38.9 % of participants exhibited a numerically positive interleaving effect, 37.7 % exhibited a numerically negative interleaving effect, and 23.5 % exhibited no interleaving effect. Across the sample, interleaving effect magnitude ranged from −0.78 to 0.67 proportion correct.

2.2.2. Fluid intelligence, episodic memory ability, and the interleaving effect

2.2.2.1. Perceptual categories. We first report the results of linear mixed-effects models and then that of supplementary quartile-based analyses. Interaction effect results are detailed in Table 3 and scatterplots corresponding to gF and EM ability, respectively, are presented on the left-side panels of Fig. 3. Violin plots and line graphs corresponding to the quartile-based analyses are presented in the right-side panels of Fig. 3.

2.2.2.1.1. Linear mixed-effects models. Separate linear mixed-effects models were conducted using the composite scores of gF and EM. The results showed that gF scores significantly predicted classification test scores ($b = 0.033$, $SE = 0.016$, $p = .039$, $d = 0.27$), and EM scores also had a significant effect ($b = 0.078$, $SE = 0.014$, $p < .001$, $d = 0.68$). Higher composite scores for both gF and EM were associated with increased classification test scores. A significant interaction effect was also found between gF composite scores and the learning schedule ($b = 0.037$, $SE = 0.014$, $p = .011$, $d = 0.41$). This interaction highlights that while gF scores predicted classification test scores in both the blocked condition ($b = 0.033$, *95 % CI* = [0.00, 0.06]) and the interleaved condition ($b = 0.070$, *95 % CI* = [0.04, 0.10]), the prediction slope was significantly steeper in the interleaved condition compared to the blocked condition. Further, a significant interaction effect was observed between EM composite scores and the learning schedule ($b = 0.041$, $SE = 0.014$, $p = .004$, $d = 0.46$). Specifically, EM scores significantly and positively predicted classification test scores in both the blocked condition ($b = 0.078$, *95 % CI* = [0.05, 0.11]) and the interleaved condition ($b = 0.12$, *95 % CI* = [0.09, 0.15]), with a steeper slope for the interleaved condition. Together with inspection of the scatterplots in Fig. 3, these results suggest that individuals with higher gF and EM ability scores exhibited a larger magnitude interleaving effect, due largely to higher performance in the interleaved condition relative to lower ability participants.

Table 3
Moderating effects of individual differences in fluid intelligence and episodic memory ability on the interleaving effect in linear mixed-effect models for Study 1.

Stimulus type	Cognitive ability	*B*	*SE*	*p*-value	Cohen's *d*
Perceptual categories					
	Fluid intelligence	0.037	0.014	0.011*	0.41
	Episodic memory	0.041	0.014	0.004**	0.46
Text-based categories					
	Fluid intelligence	−0.00072	0.019	0.969	−0.0061
	Episodic memory	−0.0015	0.019	0.937	−0.013

Note. * = $p < .05$; ** = $p < .01$. Analyses were performed using composite *z*-scores for each cognitive ability.

2.2.2.1.2. Quartile-based analyses. We performed supplementary analyses involving the lowest and highest quartiles of each composite *z*-score for gF and EM ability, respectively. Mixed-factors ANOVAs revealed a significant interaction for gF, $F(80) = 6.95$, $p = .010$, $\eta_p^2 = 0.080$, and EM ability, $F(80) = 4.10$, $p = .046$, $\eta_p^2 = 0.049$, which indicates that the interleaving effect was larger for participants with higher gF and/or higher EM ability composite scores. That pattern, which is consistent with the findings from the linear mixed-effects models, is also evident upon examination of the quartile graphs in Fig. 3.

2.2.2.2. Text-based categories. Although there was not a significant interleaving effect for text-based categories across the entire sample, we

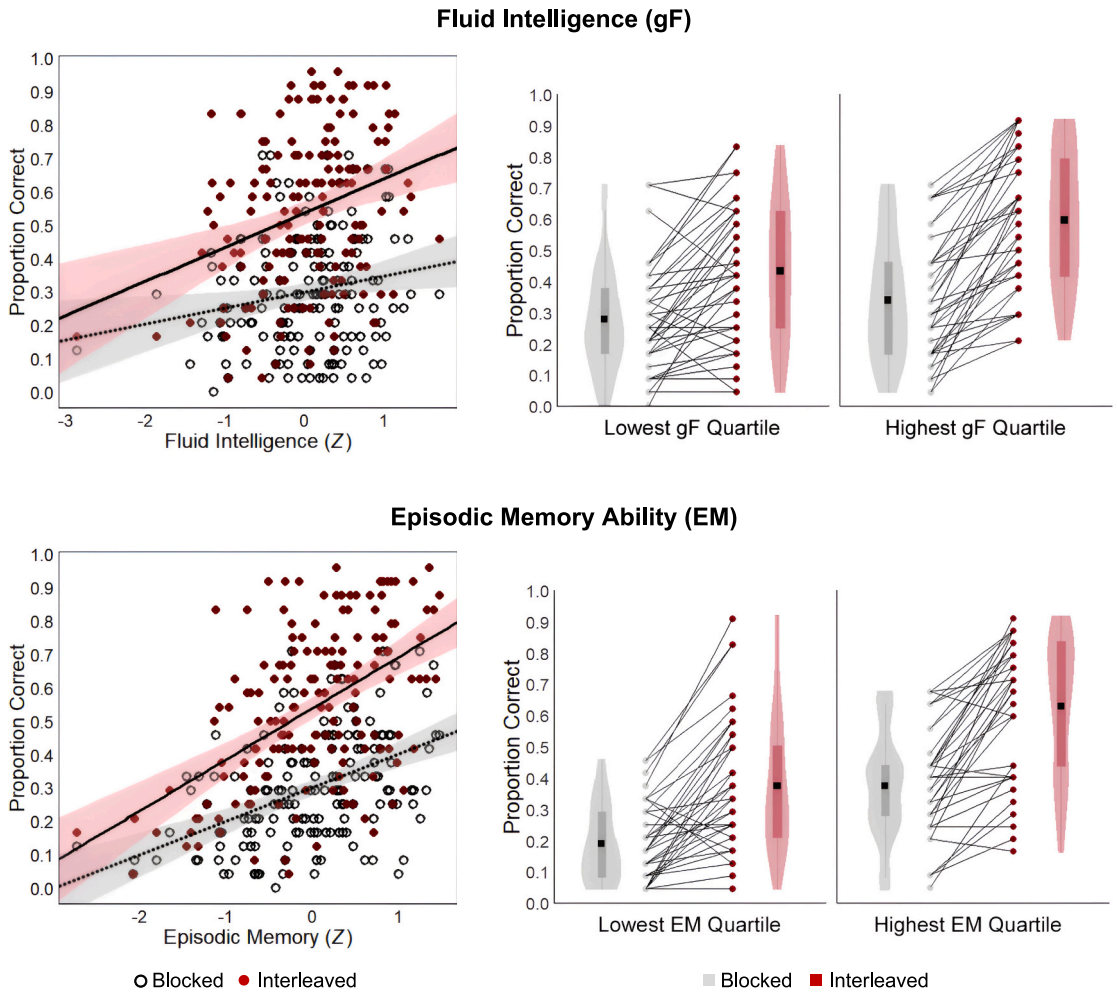


Fig. 3. Perceptual category classification test performance as a function of fluid intelligence and episodic memory ability composite scores in Study 1.
Note: In the left side panels, lines = best fitting regression line (solid and dotted lines refer to the interleaved and blocked conditions, respectively) and shading = 95 % CI. In the right side panels, the violin plots represent data from the blocked and interleaved conditions, respectively, on the visual classification test; the dot-and-line graphs represent performance in the blocked and interleaved conditions for individual participants.

still performed linear mixed-effects models to explore the potential role of individual differences in gF and EM ability. Interaction effects results are detailed in Table 3. The analyses showed that gF and EM composite z-scores positively and significantly predicted classification test scores (p-values $< .001$); that is, higher gF and EM scores were both associated with higher classification test scores. No significant interactions, however, were found between learning schedule and gF ($p = .969$) or EM ability ($p = .937$) composite scores. Given the lack of any significant interactions, no further analyses involving text-based categories were performed.

2.2.3. *Working memory capacity and the interleaving effect*

A linear mixed-effects analysis using operation span z-scores and perceptual category test scores found no significant interaction with learning schedule ($p = .544$), suggesting that WMC does not moderate the interleaving effect for perceptual category learning. A corresponding analysis for text-based categories also found no significant interaction with learning schedule ($p = .933$).

3. Study 2

The first study found that individual differences in gF and EM ability, but not WMC, moderated the effects of interleaved practice for perceptual category learning. With respect to text-based categories, however, a significant overall interleaving effect was not observed, and moreover, there were no signs that the assessed individual differences in cognitive abilities played a moderating role. To enhance the robustness of our findings and explore the generalizability of these effects, a second study was conducted to determine whether similar patterns would be observed when the same tasks and ability measures are administered to a different sample that is drawn from different population in geographical regions that were not represented in Study 1. It closely mirrored the first study in most aspects, with the primary difference being that it was conducted entirely online and sampled a completely different population that did not include undergraduate students from Singapore as in Study 1. This shift in sampling may allow for a broader range of cognitive abilities in Study 2, which could enhance the ability to detect nuanced relationships between cognitive abilities and the interleaving effect (for related discussions, see Pan et al., 2015; Unsworth, 2019).

3.1. *Methods*

The study design, research questions, and analysis plan were preregistered at https://aspredicted.org/WTR_ZGK. Nearly all aspects of Study 2, excepting sample characteristics and the online setting, were identical to that of Study 1.

3.1.1. Participants

Participants were recruited online from Prolific Academic (Prolific, London, UK), a crowdsourcing platform that is widely used in academic research and is known for its data quality (Palan & Schitter, 2017). Each participant received at least US$14.00 in exchange for their participation. All participants had to be residing in an English-speaking country, possess fluency in English, and fall within the age range of 21 to 45 years (the lower limit determined by ethics board requirements for online studies). They also had to have an approval rate of 95 % or higher on prior Prolific studies.

We again aimed for a sufficient number of participants to exceed the 120-participant minimum sample size target. Sampling occurred over two rounds of data collection and all participants met the aforementioned screening requirements. The first round involved 30 Prolific participants that had previously been recruited for a separate, one-session study in which they completed all of the same ability measures as well as an unrelated paired associate word learning task; those participants were then re-invited to complete both interleaving tasks in a follow-up 30-min session (which established that those tasks could be feasibly conducted online). In the second round, 127 entirely new Prolific participants completed the entire study in a single session (Note: despite completing the interleaving tasks in a separate session, first round participants showed no notable performance differences from second round participants). After data from one participant in the first round was excluded for technical reasons, all remaining participants from both rounds were combined into a single dataset for analysis.

The final sample of 156 participants had a mean age of 32.5 years (range: 21 to 45 years) and was 59.6 % male; in terms of ethnic background, 51.6 % were White, 22.3 % were Black, 19.7 % were Asian, and 5.7 % identified as Mixed or of other ethnic groups. Nearly half (45.5 %) of participants were from the U.K., with the remainder from the U.S. (23.1 %), Canada (19.2 %), Australia (10.9 %), and New Zealand (1.3 %). In terms of educational background, 43.6 % reported having attained an undergraduate degree as their highest level of education, whereas 21.2 %, 20.5 %, and 10.3 % indicated having attained a graduate degree, high school diploma, or other education levels, respectively.

3.1.2. Materials, procedure, data processing, and analysis plan

Nearly all aspects of this study were identical to its predecessor, including the use of the same materials, procedures, and data processing and analysis steps. The chief exceptions reflected the online setting. Specifically, all participants completed the study remotely at a time and place of their choosing. The study instructions advised them to choose a quiet, undisturbed location with a stable internet connection. We also required the use of a desktop or laptop computer that was equipped with the Google Chrome browser and the Inquisit Web 5 application, plus implemented a technical check for each participant to verify hardware and software compatibility. Finally, to ensure that participants remained focused on the task at hand, their browser activity was monitored throughout the study using TaskMaster (Permut, Fisher, & Oppenheimer, 2019). No participants were excluded from the study based on TaskMaster data, indicating satisfactory compliance with instructions.

The preregistered analysis plan specified the same sets of analyses that were performed for Study 1 with one addition: supplementary analyses of interleaving effect magnitude among participants that did or did not have an undergraduate degree (these analyses are reported in the Supplementary Results).

3.2. Results

Descriptive statistics and split-half reliability values are presented in Table 4 and a correlation matrix of all measures is presented in Table 5. It is notable that the reliability of the gF and WMC measures was numerically higher than that for Study 1, and moreover, the range of ability scores was generally wider for measures of gF, EM ability, and WMC in Study 2 versus Study 1. The following analyses are presented in the same order as in Study 1.

Table 4
Descriptive statistics for Study 2.

Measure	Mean	*SD*	Skewness	Kurtosis	Reliability
Perceptual categories					
Interleaved condition	0.48	0.24	0.066	−0.75	0.88
Blocked condition	0.30	0.19	0.92	0.57	0.86
Interleaving effect	0.18	0.19	0.098	−0.43	0.62
Text-based categories					
Interleaved condition	0.54	0.29	−0.086	−1.20	0.75–0.79[a]
Blocked condition	0.49	0.27	0.23	−0.92	0.68–0.76[a]
Interleaving effect	0.049	0.26	0.32	0.075	0.35–0.44[a]
Fluid intelligence (gF)					
Raven's matrices	0.61	0.26	−0.27	−1.11	0.91
Letter sets	0.47	0.14	−0.096	−0.38	0.72
Number series	0.56	0.19	−0.16	−0.14	0.72
Episodic memory (EM) ability					
Delayed free recall	0.50	0.20	0.47	−0.60	0.93
Cued recall	0.52	0.25	0.12	−0.91	0.92
Image recognition	0.88	0.12	−1.60	3.38	0.61
Working memory capacity (WMC)					
Operation span	58.38	17.88	−1.73	2.45	0.90

[a] Range of values reflects split-half reliabilities for the different counterbalanced versions of the text-based category learning task.

3.2.1. Interleaving effects

3.2.1.1. Perceptual categories. Participants tended to classify examples from interleaved artists more accurately than examples from blocked artists, $t(155) = 11.88$, $p < .0001$, $d = 0.95$, $BF_{10} > 1000$. The mean interleaving effect for perceptual categories was 0.18 proportion correct, which is comparable to prior findings in the literature. Variability in the effect was observed: 78.8 % of participants exhibited a numerically positive interleaving effect, 15.4 % exhibited a numerically negative interleaving effect, and 5.7 % exhibited no interleaving effect. Across the entire sample, interleaving effect magnitude ranged from −0.33 to 0.62.

3.2.1.2. Text-based categories. Participants tended to classify examples describing interleaved disorders more accurately than examples describing blocked disorders, $t(155) = 2.41$, $p = .017$, $d = 0.19$, $BF_{10} = 1.46$. The mean interleaving effect for text-based categories was 0.049 proportion correct. Albeit relatively modest in effect size (and with a Bayes factor that suggests mild evidence for the alternative hypothesis), that statistically significant difference differs from the results for Study 1 and is in closer alignment with findings reported by Zulkiply et al. (2012) and Pan, Selvarajan and Murphy (2024). Variability in that effect was also observed: 45.5 % of participants exhibited a numerically positive interleaving effect, 36.5 % exhibited a numerically negative interleaving effect, and 17.9 % exhibited no interleaving effect. Across the entire sample, interleaving effect magnitude ranged from −0.55 to 0.78.

3.2.2. Fluid intelligence, episodic memory ability, and the interleaving effect

3.2.2.1. Perceptual categories. As with Study 1, we first report the results of linear-mixed effects models involving gF and EM ability, followed by quartile-based analyses. Interaction effect results are detailed in Table 6 and scatterplots corresponding to gF and EM ability, respectively, are presented on the left-side panels of Fig. 4. Violin plots and line graphs corresponding to the quartile-based analyses are presented in the right-side panels of Fig. 5.

3.2.2.1.1. Linear mixed-effects models. We submitted classification test scores to linear mixed-effects models involving composite *z*-scores of

Table 5
Correlation matrix for Study 2.

Measure	1	2	3	4	5	6	7	8	9	10	11	12	13
Interleaving tasks													
1. I-Visual	1.00												
2. B-Visual	0.63***	1.00											
3. IE-Visual	0.61***	−0.22**	1.00										
4. I-Text	0.42***	0.26***	0.26**	1.00									
5. B-Text	0.40***	0.28***	0.22**	0.58***	1.00								
6. IE-Text	0.048	0.0022	0.058	0.52***	−0.39***	1.00							
gF tasks													
7. Raven	0.40***	0.28***	0.22**	0.32***	0.25**	0.099	1.00						
8. Lsets	0.19*	0.12	0.11	0.23**	0.11	0.15	0.38***	1.00					
9. Nseries	0.31***	0.12	0.26***	0.30***	0.26**	0.069	0.49***	0.49***	1.00				
EM ability tasks													
10. DFR	0.21**	0.26**	−0.0039	0.15	0.21**	−0.049	0.22**	−0.0077	0.064	1.00			
11. CR	0.25**	0.28***	0.034	0.12	0.23**	−0.11	0.13	0.10	0.18*	0.61***	1.00		
12. Recog	0.33***	0.17*	0.24**	0.21**	0.18*	0.048	0.31***	0.073	0.12	0.22**	0.29***	1.00	
WMC task													
13. Ospan	0.27***	0.15	0.19*	0.16*	0.14	0.037	0.23**	0.14	0.26***	0.37***	0.39***	0.22**	1.00

Note. * = $p < .05$; ** = $p < .01$; *** = $p < .001$. For the interleaving tasks (1–6): I = interleaved condition; B = blocked condition, IE = interleaving effect; Visual = perceptual categories; Text = text-based categories. For the fluid intelligence tasks (7–9): Raven = Raven's progressive matrices; Lsets = Letter sets; Nseries = Number series. For the episodic memory tasks (10–12): DFR = delayed free recall; CR = cued recall; Recog. = image recognition. Working memory task (13): Ospan = operation span.

Table 6
Moderating effects of individual differences in fluid intelligence and episodic memory ability on the interleaving effect in linear mixed-effect models for Study 2.

Stimulus type	Cognitive ability	*B*	*SE*	*p*-value	Cohen's *d*
Perceptual categories					
	Fluid intelligence	0.047	0.015	0.002**	0.51
	Episodic memory	0.023	0.015	0.136	0.24
Text-based categories					
	Fluid intelligence	0.034	0.020	0.101	0.27
	Episodic memory	−0.012	0.020	0.565	−0.093

Note. * = $p < .05$; ** = $p < .01$. Analyses were performed using composite *z*-scores for each cognitive ability.

gF and EM ability. In these analyses, gF scores significantly predicted classification test scores, $b = 0.043$, $SE = 0.017$, $p = .011$, $d = 0.34$, as did EM scores, $b = 0.059$, $SE = 0.017$, $p < .001$, $d = 0.47$. There was a significant interaction effect between gF composite scores and learning schedule, $b = 0.047$, $SE = 0.015$, $p = .002$, $d = 0.51$. Specifically, gF composite scores significantly and positively predicted classification test scores in both the blocked condition ($b = 0.043$, 95 % CI = [0.01, 0.08]) and the interleaved condition ($b = 0.090$, 95 % CI = [0.06, 0.12]), with a steeper slope for the interleaved condition. No significant interaction was found between learning schedule and composite scores for EM ability ($p = .136$). These results are depicted in the scatterplots of Fig. 4.

3.2.2.1.2. Quartile-based analyses. We performed supplementary analyses involving the lowest and highest quartiles of the composite *z*-scores for gF and EM ability. Mixed-factors ANOVAs revealed a significant interaction for gF, $F(76) = 10.88$, $p = .001$, $\eta_p^2 = 0.125$, but not for EM ability ($p = .147$). That pattern, which aligns with the linear mixed-effects model findings, is evident upon examination of the quartile graphs in Fig. 4, in which the highest gF quartile exhibited a larger interleaving effect due to higher performance in the interleaved condition.

3.2.2.2. Text-based categories. Linear mixed-effects analyses both showed that gF and EM composite scores positively and significantly predicted classification test scores (*p*-values < .003). No significant interactions were found between learning schedule and composite scores for gF ($p = .101$) or EM ability ($p = .565$). Given these results, we did not perform any further analyses involving text-based categories.

3.2.3. Working memory capacity and the interleaving effect

A linear mixed-effects analysis involving operation span *z*-scores and visual classification test performance revealed a significant interaction with learning schedule, $b = 0.036$, $SE = 0.015$, $p = .017$, $d = 0.39$. Specifically, operation span scores significantly and positively predicted classification test scores in the interleaved condition ($b = 0.065$, 95 % CI = [0.03, 0.10]), whereas the relationship was not significant in the blocked condition ($b = 0.029$, 95 % CI = [−0.00, 0.06]). This result contrasts with findings from Study 1 and other reports in the literature (e.g., Sana et al., 2017, 2018; Wang et al., 2020) and is more reminiscent of the patterns reported by Guzman-Munoz (2017). A corresponding analysis involving text-based categories found no significant interaction between learning schedule and operation span *z*-scores ($p = .649$), similar to Study 1.

4. Structural equation modeling analyses of studies 1 and 2 combined

To examine how gF, EM ability, and WMC might influence the interleaving effect when all three abilities are considered simultaneously, we employed a structural equation modeling (SEM) approach. SEM allows for the simultaneous examination of multiple relationships between observed and latent variables (Hair et al., 2006). To meet the sample size requirement of at least 200 participants for SEM, we combined the datasets from both studies (combined $n = 318$). SEM models were fitted using the lavaan package (version 0.6–9; Rosseel, 2012) in R, with robust maximum likelihood estimation and robust standard errors. The SEM analyses were not preregistered.

4.1. Perceptual categories

In the baseline model (Model 1), which is illustrated in Fig. 1, we used the task scores corresponding to each cognitive ability as indicators for the latent variables: gF, EM ability, and WMC. We fixed the variance of all latent factors to one and regressed these three latent variables on the classification test scores from both the interleaved and blocked conditions. Additionally, we allowed covariances between the classification test scores in both conditions and among the latent factors. The model fit indices were: $\chi^2(20) = 64.05$, $p < .001$, SRMR = 0.053, robust CFI = 0.93, robust RMSEA = 0.083, 90 % CI [0.061, 0.107]. RMSEA values between 0.05 and 0.08 suggest a reasonable fit and values above 0.10 indicate a poor fit (MacCallum et al., 1996), whereas SRMR values below 0.08 and CFI values above 0.90 are considered indicative of an

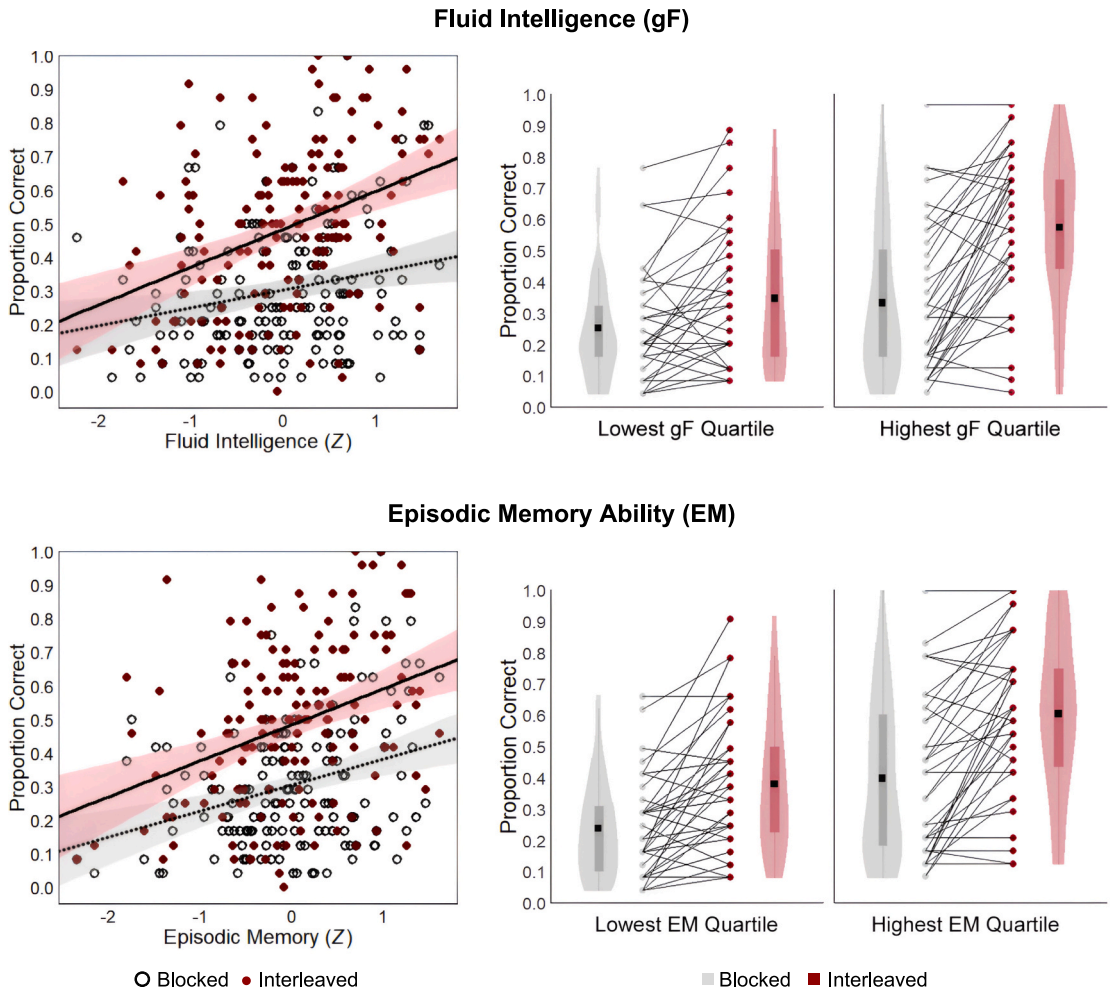


Fig. 4. Perceptual category classification test performance as a function of fluid intelligence and episodic memory ability composite scores in Study 2.
Note: In the left side panels, lines = best fitting regression line (solid and dotted lines refer to the interleaved and blocked conditions, respectively) and shading = 95 % CI. In the right side panels, the violin plots represent data from the blocked and interleaved conditions, respectively, on the visual classification test; the dot-and-line graphs represent performance in the blocked and interleaved conditions for individual participants.

acceptable fit (Browne & Cudeck, 1992; Hu & Bentler, 1999). Hence, although the RMSEA value suggests a mediocre fit, the other indices indicate an acceptable model fit.

To assess whether the regression slopes for classification test scores differed between the interleaved and blocked conditions for each cognitive ability, we fitted constrained models. Specifically, we developed three constrained models (Models 2, 3, and 4), each examining the equivalence of regression slopes for one cognitive ability. In Model 2, the regression slopes from gF to the classification test scores were constrained to be equal across conditions. Model 3 applied the same constraint for EM ability, while Model 4 did so for WMC.

The fit of Model 2, which constrained the slopes for gF, was significantly worse than that of Model 1, $\Delta\chi^2(1) = 10.40$, $p = .00126$. This finding suggests that the regression slopes from gF to classification test scores differed significantly between the interleaved and blocked conditions. Specifically, gF significantly predicted test scores in the interleaved condition ($b = 0.074$, $SE = 0.017$, $\beta = 0.31$, $p < .001$), while it did not significantly predict test scores in the blocked condition ($b = 0.017$, $SE = 0.015$, $\beta = 0.093$, $p = .251$). Thus, gF appears to moderate the interleaving effect by influencing performance in the interleaved condition—a finding that mirrors the analyses performed separately for Studies 1 and 2. The fit of Model 3 (with constrained EM ability slopes) did not significantly differ from that of Model 1, $\Delta\chi^2(1) = 0.065$, $p = .799$, nor did the fit of Model 4 (with constrained WMC slopes), $\Delta\chi^2(1) = 1.35$, $p = .245$.

4.2. Text-based categories

A corresponding SEM analysis was conducted involving the text-based categories. We fitted one baseline model (Model 5) and three constrained models (Models 6, 7, and 8). The baseline model demonstrated an acceptable fit, with the following indices: $\chi^2(20) = 55.93$, $p < .001$, SRMR = 0.053, robust CFI = 0.94, robust RMSEA = 0.075, 90 % CI [0.051, 0.099]. Model 6, which constrained the slopes for gF, did not significantly differ from that of Model 5, $\Delta\chi^2(1) = 2.22$, $p = .136$, and the same was true for Model 7, which constrained the slopes for EM ability, $\Delta\chi^2(1) = 2.03$, $p = .154$, and Model 8, which constrained the slopes for WMC, $\Delta\chi^2(1) = 0.0025$, $p = .960$. Overall, these findings suggest that none of the investigated cognitive abilities moderated the interleaving effect for text-based categories.

5. Discussion

The present investigation found that individual differences in cognitive abilities moderate the interleaving effect for perceptual category learning. Specifically, in both studies, individuals with higher gF scores exhibited a larger magnitude interleaving effect for learning landscape artists' painting styles. Linear mixed-effects models and

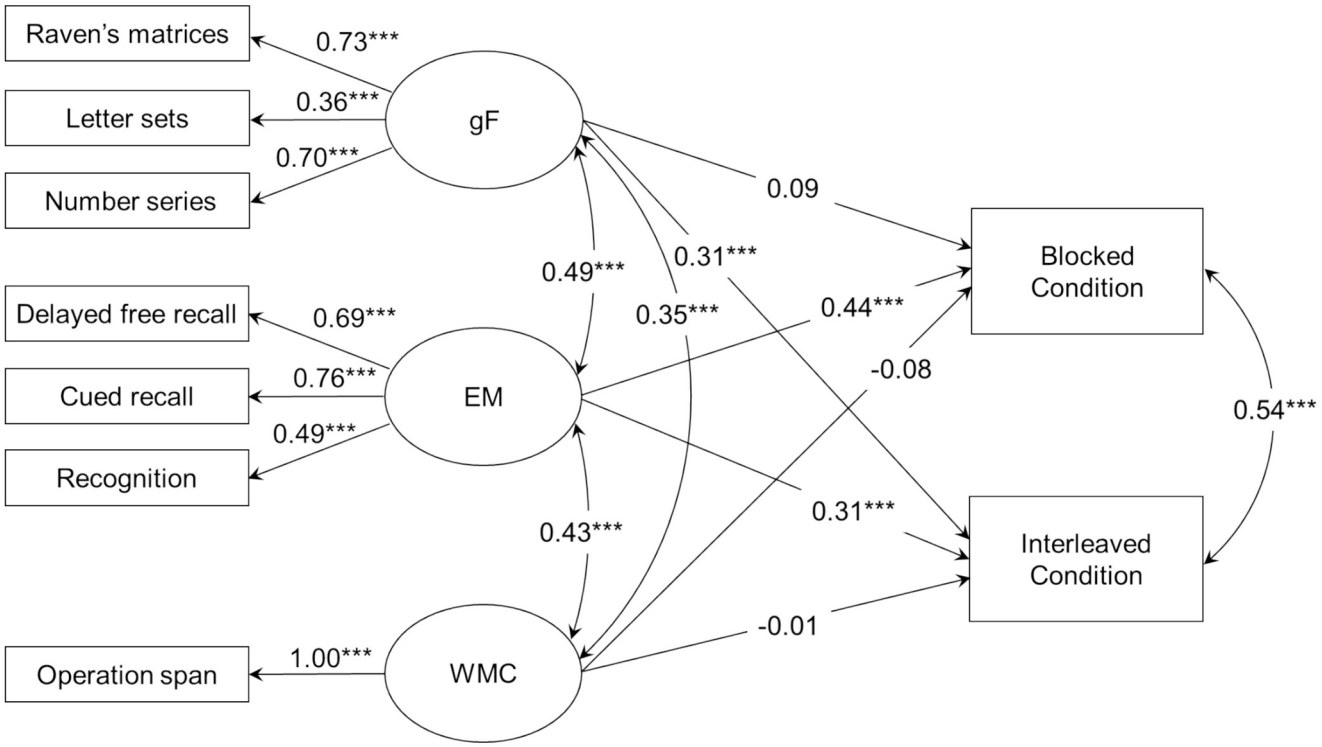


Fig. 5. Structural equation model for fluid intelligence, episodic memory ability, working memory capacity and perceptual category classification test performance in Studies 1 and 2.
Note. gf = fluid intelligence; EM = episodic memory ability; WMC = working memory capacity. *** = $p < .001$. Baseline model shown.

quartile-based analyses, as well as structural equation modeling analyses of combined data from both studies, all yielded results consistent with that pattern. Similar patterns were observed for individuals with higher EM ability scores in Study 1 and higher WMC scores in Study 2. Further, supplementary analyses conducted using factor scores yielded the same patterns as the analyses with composite *z*-scores (see Supplementary Results).

Overall, across all three investigated cognitive abilities, there were indications that higher-ability individuals benefited more from interleaving (although, at least for perceptual categories, there were at least some benefits across the measured ability ranges). No moderating effects of cognitive abilities, however, were found for the interleaving effect involving text-based category learning. Thus, with respect to interleaving and perceptual category learning, our results support the first possibility outlined earlier in this manuscript (as illustrated in the leftmost panel of Fig. 2): Higher-ability learners benefit the most.

5.1. Fluid intelligence, episodic memory ability, and the interleaving effect

Our most prominent finding, consistently observed across two samples varying in age, educational background, geographical setting, and other demographics, centered on individual differences in gF. In both studies, higher-gF individuals showed a more pronounced interleaving effect, primarily due to better performance in the interleaved condition compared to their lower-gF counterparts. Hence, for perceptual category learning, higher-gF individuals seem better able to capitalize on the learning opportunities that interleaving provides. One possibility is that their advanced abstract reasoning abilities (Carroll, 1993) enhance discriminative contrast processes. Another possibility involves viewing interleaved sequences as problem-solving scenarios where the object is to discover patterns that differentiate categories. The repeated juxtaposition of different categories in interleaving allows for constant hypothesis testing, which blocking does not facilitate. In these scenarios, higher-gF individuals may excel at generating viable solutions (see also Greiff & Neubert, 2014; P. Kyllonen et al., 2017; cf. Little & McDaniel, 2015).

As for the finding that higher-EM ability individuals showed a larger interleaving effect for perceptual categories in Study 1, a possible explanation is that these individuals demonstrated improved recall of features unique to previously studied artists from long-term memory. This enhancement may be due in part to their greater engagement in study-phase retrieval, which in turn bolstered the effectiveness of interleaving for acquiring classification skills. In contrast, lower-gF and lower-EM ability individuals may have had greater difficulties recalling category features or other information, leading to less learning.

A related possibility involves the greater use of effective memory strategies among higher-EM ability and/or higher-gF individuals (e.g., Kirchhoff, 2009; Minear et al., 2018). Such strategies may be inherently more conducive to interleaving (as opposed to those strategies obviating the need to engage in interleaving to promote learning). Strategy use may also enable higher-ability individuals to better remember the between-category distinctions that they learned through interleaving. If so, then memory strategy use may be another viable explanation for the greater benefits of interleaving for higher-EM or higher-gF individuals.

Adjudicating between all of these possible explanations for the observed patterns will require further research. Moreover, a caveat to the foregoing discussion is that while gF consistently showed a moderating role across both studies, the same was not observed for EM ability. Although the patterns for EM ability in Study 2 arguably did not diverge dramatically from those observed in Study 1, it remains unclear whether that inconsistency originated from different sample characteristics or other factors.

5.2. Working memory capacity and the interleaving effect

The present results present a mixed picture regarding WMC. In Study 1, the absence of a moderating role for WMC aligns with findings by Sana et al. (2018), Wang et al. (2020), and Yan and Sana (2021), who

used similar learning materials and primarily sampled undergraduate students (excepting Sana et al., Experiments 2b and 4). Contrasting with those studies are the results of Study 2, in which there was a larger interleaving effect for higher-WMC individuals (driven by better performance in the interleaved condition, similar to observed patterns for gF). Those findings echo Guzman-Munoz (2017; Experiments 2 and 3), which reported similar but marginally significant results derived from smaller sample sizes.

The failure to find a moderating role of WMC on the interleaving effect may stem from various scenarios. First, interleaving may simply benefit most learners regardless of WMC. Relatedly, the interleaving effect for perceptual categories may rely on cognitive processes and systems that are not as heavily impacted by WMC (e.g., non-declarative learning). Another possibility is that different processes are predominant during interleaving in high- versus low-WMC individuals (possibly a spacing effect in the former and discriminative contrast in the latter), yielding comparable effects (Sana et al., 2018; an account involving different mechanisms for different ability learners might also be applicable to gF or EM ability).

On the other hand, there may be as-yet-unclear circumstances where a moderating role of WMC may emerge (as in Study 2). Such circumstances might include a broader range of WMC abilities (as was the case in Study 2 versus Study 1), or in specific contexts, such as different study settings or materials. Due to the inconsistent results in our investigation and prior null findings, however, further research is needed to explore the role of WMC with different populations and learning materials (for related discussion see Krefeld-Schwalb et al., 2024).

5.3. Interleaving effects for perceptual versus text-based categories

Consistent with patterns reported by Brunmair and Richter (2019), the interleaving effect for artists' painting styles was more pronounced and consistent than for psychological disorders ($ds = 1.29$ and 0.95 vs. 0.020 and 0.19 for perceptual and text-based categories, respectively, in Studies 1 and 2, with a significant interleaving effect for the latter not appearing in Study 1). Three-quarters of participants in both studies showed at least a numerically positive interleaving effect for perceptual categories, while less than half did for text-based categories. These patterns suggest that the interleaving effect is generally weaker for text-based materials, although exceptions exist (e.g., Abel et al., 2021). This finding aligns with prior research (e.g., Yan & Sana, 2021) indicating weak correlations between interleaving effects across different stimulus classes (see Tables 2 and 5). Additionally, the limited number of categories learned—6 for text-based versus 12 for perceptual categories—may contribute to a smaller interleaving effect, although interleaving has been effective even with as few as two categories (e.g., Pan et al., 2025; see also Schweppe, Lenk-Blochowitz, Pucher, & Ketzer-Nöltge, 2024)

The lack of a significant interleaving effect for psychological disorders in Study 1 is not unprecedented. Murphy and Pavlik (2018) attempted a near-direct replication of Zulkiply et al. (2012), substituting common names for psychological disorders instead of nonsense names, and also found no interleaving effect. As previously noted, however, the case study materials used in the present research previously yielded a significant interleaving effect in a between-subjects design (Pan, Selvarajan, & Murphy, 2024), as well as in Study 2. Another possible reason for these inconsistent results is that text-based category learning may be more influenced by reading ability and other skills than perceptual category learning. Those individual differences, along with background knowledge about psychological disorders (despite our exclusion criteria), could have impacted the interleaving effect for these materials.

5.4. Study limitations and future research

Although prior individual differences studies have employed similar sample sizes as in the present research, future studies with larger samples may add further insights (Unsworth, 2019), including with respect to the replicability of the present findings. We recommend further exploring the role of gF and EM ability (and possibly reexamining the role of WMC) with additional samples, including from different sources as in existing research. Future studies could also address other forms of individual differences, for instance in the amount of perceived or actual effort (Abel, De Bruin, et al., 2024; Onan et al., 2022), as well as differences in cognitive processes or performance among different ability learners when blocked schedules are used.

This investigation was limited to two types of stimulus materials. Future research should explore the interleaving effect across other, more widely-learned materials (for discussion see Pan, González-Cabañes, et al., 2024), such as problem-solving skills (e.g., Rohrer et al., 2015) and language learning (e.g., Suzuki et al., 2022). It is possible that the moderating effects of cognitive abilities on the interleaving effect for one class of materials may be different with other stimulus materials. Such studies might even employ longer test delays to address the interleaving effect with more educationally-relevant retention intervals than in the present research.

Finally, future studies could investigate whether individual differences in academic performance (and not just attainment) moderate the benefits of interleaving. Given that the interleaving effect was larger among individuals with higher gF and EM ability scores, it is possible that stronger academic performers, who often exhibit higher levels of these cognitive abilities (Alloway & Alloway, 2010; Blankenship et al., 2015; Colom et al., 2007; Di Fabio & Busoni, 2007; Ren et al., 2015), may benefit more from interleaving.

5.5. Pedagogical implications

The finding that interleaving especially benefits higher-ability learners for perceptual category learning carries significant implications for educational practice. Specifically, it suggests that interleaving does not uniformly help all learners; rather, it can disproportionately benefit higher-ability learners (i.e., it helps the “rich get richer”). This pattern contrasts with the notion that most “desirable difficulties” engage atypical learning processes (McDaniel & Butler, 2011; Pan & Bjork, 2022), which would potentially result in greater benefits for less skilled learners (see also Nemeth & Lipowsky, 2023).

Despite our finding of greater interleaving benefits for higher-ability learners, it is crucial to emphasize that the vast majority of participants in the present studies, including many participants with lower gF or EM ability scores, still profited from interleaving for perceptual category learning. Thus, a general recommendation to use interleaving for such learning remains justified.

5.6. Conclusions

We found that the interleaving effect for perceptual category learning is moderated by fluid intelligence, with higher-gF individuals exhibiting a larger interleaving effect. That larger effect was driven by markedly improved performance for materials learned through interleaving among higher-ability individuals. There were also some indications that higher-EM ability and higher-WMC individuals may also exhibit a larger interleaving effect, again driven by improved performance for materials learned through interleaving for higher-ability individuals. No moderating effects of gF, EM ability, or WMC were observed regarding the interleaving effect for text-based categories. Together, these results suggest that interleaving especially benefits higher-ability learners in the case of perceptual category learning, with implications for the use of interleaving in authentic educational contexts.

CRediT authorship contribution statement

Steven C. Pan: Writing – review & editing, Writing – original draft,

Supervision, Software, Resources, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Conceptualization. **Liwen Yu:** Writing – review & editing, Writing – original draft, Formal analysis, Data curation. **Yilin Hong:** Supervision, Project administration, Methodology, Investigation, Formal analysis, Data curation. **Marcus J. Wong:** Methodology, Investigation, Data curation. **Ganeash Selvarajan:** Software, Resources, Methodology, Investigation, Data curation. **Michelle E. Kaku:** Supervision, Resources, Investigation, Data curation.

Declaration of competing interest

We have no conflicts of interest to disclose.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.lindif.2024.102603.

References

Abel, R., De Bruin, A., Onan, E., & Roelle, J. (2024). Why do learners (under)utilize interleaving in learning confusable categories? The role of metastrategic knowledge and utility value of distinguishing. *Educational Psychology Review, 36*(2), 64. https://doi.org/10.1007/s10648-024-09902-0

Abel, R., Niedling, L. M., & Hänze, M. (2021). Spontaneous inferential processing while reading interleaved expository texts enables learners to discover the underlying regularities. *Applied Cognitive Psychology, 35*(1), 258–273. https://doi.org/10.1002/acp.3761

Abel, R., Roelle, J., & Stadtler, M. (2024). Whom to believe? Fostering source evaluation skills with interleaved presentation of untrustworthy and trustworthy social media sources. *Discourse Processes*. https://doi.org/10.1080/0163853X.2024.2339733

Alloway, T. P., & Alloway, R. G. (2010). Investigating the predictive roles of working memory and IQ in academic attainment. *Journal of Experimental Child Psychology, 106*(1), 20–29. https://doi.org/10.1016/j.jecp.2009.11.003

Bailey, H., Dunlosky, J., & Kane, M. J. (2008). Why does working memory span predict complex cognition? Testing the strategy affordance hypothesis. *Memory & Cognition, 36*, 1383–1390. https://doi.org/10.3758/MC.36.8.1383

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

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

Bjork, E. L., & Bjork, R. A. (2011). Making things hard on yourself, but in a good way: Creating desirable difficulties to enhance learning. In *Psychology and the real world: Essays illustrating fundamental contributions to society* (pp. 59–68).

Blankenship, T. L., O'Neill, M., Ross, A., & Bell, M. A. (2015). Working memory and recollection contribute to academic achievement. *Learning and Individual Differences, 43*, 164–169. https://doi.org/10.1016/j.lindif.2015.08.020

Brewer, G. A., & Unsworth, N. (2012). Individual differences in the effects of retrieval from long-term memory. *Journal of Memory and Language, 66*(3), 407–415. https://doi.org/10.1016/j.jml.2011.12.009

Browne, M. W., & Cudeck, R. (1992). Alternative ways of assessing model fit. *Sociological Methods & Research, 21*(2), 230–258. https://doi.org/10.1177/0049124192021002005

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

Carpenter, S. K. (2014). Spacing and interleaving of study and practice. In V. A. Benassi, C. E. Overson, & C. M. Hakala (Eds.), *Applying the science of learning in education: Infusing psychological science into the curriculum*. American Psychological Association. http://teachpsych.org/resources/documents/ebooks/asle2014.pdf#page=137.

Carpenter, S. K., & Pan, S. C. (2024). Spacing effects in learning and memory. In *Reference module in neuroscience and biobehavioral psychology*. Elsevier, Article B9780443157547000201. https://doi.org/10.1016/B978-0-443-15754-7.00020-1.

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

Carroll, J. B. (1993). *Human cognitive abilities: A survey of factor-analytic studies*. Cambridge University Press.

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

Carvalho, P. F., & Goldstone, R. L. (2019). When does interleaving practice improve learning? In J. Dunlosky, & K. A. Rawson (Eds.), *The Cambridge handbook of cognition and education* ((1st ed., pp. 411–436). Cambridge University Press. https://doi.org/10.1017/9781108235631.017.

Cepeda, N. J., Pashler, H., Vul, E., Wixted, J. T., & Rohrer, D. (2006). Distributed practice in verbal recall tasks: A review and quantitative synthesis. *Psychological Bulletin, 132*(3), 354–380. https://doi.org/10.1037/0033-2909.132.3.354

Cleary, A. M. (2018). Dependent measures in memory research. In H. Otani, & B. L. Schwartz (Eds.), *Handbook of research methods in human memory* ((1st ed., pp. 19–35). Routledge. https://doi.org/10.4324/9780429439957-2.

Colom, R., Escorial, S., Shih, P. C., & Privado, J. (2007). Fluid intelligence, memory span, and temperament difficulties predict academic performance of young adolescents. *Personality and Individual Differences, 42*(8), 1503–1514. https://doi.org/10.1016/j.paid.2006.10.023

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*(5), 769–786. https://doi.org/10.3758/BF03196772

Del Missier, F., Sassano, A., Coni, V., Salomonsson, M., & Mäntylä, T. (2018). Blocked vs. Interleaved presentation and proactive interference in episodic memory. *Memory, 26*(5), 697–711. https://doi.org/10.1080/09658211.2017.1402937

Delaney, P. F., Verkoeijen, P. P. J. L., & Spirgel, A. (2010). Spacing and testing effects. In *, 53. Psychology of learning and motivation* (pp. 63–147). Elsevier. https://doi.org/10.1016/S0079-7421(10)53003-2.

Di Fabio, A., & Busoni, L. (2007). Fluid intelligence, personality traits and scholastic success: Empirical evidence in a sample of Italian high school students. *Personality and Individual Differences, 43*(8), 2095–2104. https://doi.org/10.1016/j.paid.2007.06.025

Ebbinghaus, H. (1885). *Über das Gedächtnis: Untersuchungen zur experimentellen Psychologie*. Duncker & Humblot.

Eglington, L. G., & Kang, S. H. K. (2017). Interleaved presentation benefits science category learning. *Journal of Applied Research in Memory and Cognition, 6*(4), 475–485. https://doi.org/10.1016/j.jarmac.2017.07.005

Ekstrom, R. B., French, J. W., & Harman, H. H. (1976). *Kit of factor-referenced cognitive tests*. Educational Testing Service.

Engle, R. W., & Kane, M. J. (2003). Executive attention, working memory capacity, and a two-factor theory of cognitive control. In *, 44. Psychology of learning and motivation* (pp. 145–199). Elsevier. https://doi.org/10.1016/S0079-7421(03)44005-X.

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

Foster, N. L., Mueller, M. L., Was, C., Rawson, K. A., & Dunlosky, J. (2019). Why does interleaving improve math learning? The contributions of discriminative contrast and distributed practice. *Memory & Cognition, 47*(6), 1088–1101. https://doi.org/10.3758/s13421-019-00918-4

Greiff, S., & Neubert, J. C. (2014). On the relation of complex problem solving, personality, fluid intelligence, and academic achievement. *Learning and Individual Differences, 36*, 37–48. https://doi.org/10.1016/j.lindif.2014.08.003

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

Hair, J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2006). *Multivariate data analysis* ((6th ed).). Pearson Prentice Hall.

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, 1*(1), 31–40. https://doi.org/10.1037/0278-7393.1.1.31

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal, 6*(1), 1–55. https://doi.org/10.1080/10705519909540118

Kachouri, H., Fay, S., Angel, L., & Isingrini, M. (2022). Influence of current physical exercise on the relationship between aging and episodic memory and fluid intelligence. *Acta Psychologica, 227*, Article 103609. https://doi.org/10.1016/j.actpsy.2022.103609

Kang, S. H. K. (2017). The benefits of interleaved practice for learning. In *From the laboratory to the classroom: Translating science of learning for teachers*. Routledge.

Kang, S. H. K., & Pashler, H. (2012). Learning painting styles: Spacing is advantageous when it promotes discriminative contrast: Spacing promotes contrast. *Applied Cognitive Psychology, 26*(1), 97–103. https://doi.org/10.1002/acp.1801

Kievit, R. A., Davis, S. W., Griffiths, J., Correia, M. M., & Cam-CAN, & Henson, R. N.. (2016). A watershed model of individual differences in fluid intelligence. *Neuropsychologia, 91*, 186–198. https://doi.org/10.1016/j.neuropsychologia.2016.08.008

Kirchhoff, B. A. (2009). Individual differences in episodic memory: The role of self-initiated encoding strategies. *The Neuroscientist, 15*(2), 166–179. https://doi.org/10.1177/1073858408329507

Kornell, N., & Bjork, R. A. (2008). Learning concepts and categories: Is spacing the "enemy of induction"? *Psychological Science, 19*(6), 585–592. https://doi.org/10.1111/j.1467-9280.2008.02127.x

Krefeld-Schwalb, A., Sugerman, E. R., & Johnson, E. J. (2024). Exposing omitted moderators: Explaining why effect sizes differ in the social sciences. *Proceedings of the National Academy of Sciences, 121*(12), Article e2306281121. https://doi.org/10.1073/pnas.2306281121

Kyllonen, P., Anguiano Carrasco, C., & Kell, H. J. (2017). Fluid ability (Gf) and complex problem solving (CPS). *Journal of Intelligence, 5*(3). https://doi.org/10.3390/jintelligence5030028. Article 3.

Kyllonen, P. C., & Christal, R. E. (1990). Reasoning ability is (little more than) working-memory capacity?! *Intelligence, 14*(4), 389–433. https://doi.org/10.1016/S0160-2896(05)80012-1

Little, J. L., & McDaniel, M. A. (2015). Individual differences in category learning: Memorization versus rule abstraction. *Memory & Cognition, 43*(2), 283–297. https://doi.org/10.3758/s13421-014-0475-1

MacCallum, R. C., Browne, M. W., & Sugawara, H. M. (1996). Power analysis and determination of sample size for covariance structure modeling. *Psychological Methods, 1*(2), 130–149. https://doi.org/10.1037/1082-989X.1.2.130

McDaniel, M. A., & Butler, A. C. (2011). A contextual framework for understanding when difficulties are desirable. In *Successful remembering and successful forgetting: A festschrift in honor of Robert A. Bjork* (pp. 175–198). Psychology Press.

Millisecond Software. (2023). *Automated Operation Span Task (AOSPAN) [Computer software]*. Retrieved from https://www.millisecond.com.

Minear, M., Coane, J. H., Boland, S. C., Cooney, L. H., & Albat, M. (2018). The benefits of retrieval practice depend on item difficulty and intelligence. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 44*(9), 1474–1486. https://doi.org/10.1037/xlm0000486

Morey, R. D., & Rouder, J. N. (2012). *BayesFactor: Computation of Bayes factors for common designs* (p. 0.9.12-4.7) [Dataset]. https://doi.org/10.32614/CRAN.package.BayesFactor.

Murphy, C. S., & Pavlik, P. I. (2018). Effects of spacing and testing on inductive learning. *Journal of Articles in Support of the Null Hypothesis, 15*(1).

Nemeth, L., & Lipowsky, F. (2023). The role of prior knowledge and need for cognition for the effectiveness of interleaved and blocked practice. *European Journal of Psychology of Education*. https://doi.org/10.1007/s10212-023-00723-3

Onan, E., Wiradhany, W., Biwer, F., Janssen, E. M., & de Bruin, A. B. H. (2022). Growing out of the experience: How subjective experiences of effort and learning influence the use of Interleaved practice. *Educational Psychology Review, 34*(4), 2451–2484. https://doi.org/10.1007/s10648-022-09692-3

Palan, S., & Schitter, C. (2017). Prolific.ac—A subject pool for online experiments. *Journal of Behavioral and Experimental Finance, 17*. https://doi.org/10.1016/j.jbef.2017.12.004

Pan, S. C., & Bjork, R. A. (2022). Acquiring an accurate mental model of human learning: Toward an owner's manual. In A. Wagner, & M. Kahana (Eds.), *Oxford handbook of learning and memory: Foundations and applications*. Oxford University Press.

Pan, S. C., Pashler, H., Potter, Z. E., & Rickard, T. C. (2015). Testing enhances learning across a range of episodic memory abilities. *Journal of Memory and Language, 83*, 53–61. https://doi.org/10.1016/j.jml.2015.04.001

Pan, S. C., Selvarajan, G., & Murphy, C. S. (2024). Interleaved pretesting enhances category learning and classification skills. *Journal of Applied Research in Memory and Cognition, 13*(3), 393–406. https://doi.org/10.1037/mac0000194

Pan, S. C., González-Cabañes, E., Teo, A., Zung, I., Sana, F., & Cooke, J. E. (2024). *Distributed practice and interleaved practice: Undergraduate students' strategies, experiences, and beliefs*. Manuscript in preparation.

Pan, S. C., Flores, S. R., Kaku, M. E., & Lai, W. H. E. (2025). Interleaved practice enhances grammar skill learning for similar and dissimilar tenses in Romance languages. *Learning and Instruction, 95*, Article 102045. https://doi.org/10.1016/j.learninstruc.2024.102045

Permut, S., Fisher, M., & Oppenheimer, D. M. (2019). TaskMaster: A tool for determining when subjects are on task. *Advances in Methods and Practices in Psychological Science, 2*(2), 188–196. https://doi.org/10.1177/2515245919838479

Raven, J., & Raven, J. (2003). Raven Progressive Matrices. In *Handbook of nonverbal assessment* (pp. 223–237). Kluwer Academic/Plenum Publishers. https://doi.org/10.1007/978-1-4615-0153-4_11.

Raykov, P. P., Knights, E., & Cam-CAN, & Henson, R. N.. (2024). Does functional system segregation mediate the effects of lifestyle on cognition in older adults? *Neurobiology of Aging, 134*, 126–134. https://doi.org/10.1016/j.neurobiolaging.2023.11.009

R Core Team. (2021). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Ren, X., Schweizer, K., Wang, T., & Xu, F. (2015). The prediction of students' academic performance with fluid intelligence in giving special consideration to the contribution of learning. *Advances in Cognitive Psychology, 11*(3), 97–105. https://doi.org/10.5709/acp-0175-z

Robey, A. (2019). The benefits of testing: Individual differences based on student factors. *Journal of Memory and Language, 108*, Article 104029. https://doi.org/10.1016/j.jml.2019.104029

Rohrer, D. (2012). Interleaving helps students distinguish among similar concepts. *Educational Psychology 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*(5), 1323–1330. https://doi.org/10.3758/s13423-014-0588-3

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

Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software, 48*, 1–36. https://doi.org/10.18637/jss.v048.i02

Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., & Iverson, G. (2009). Bayesian t tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review, 16*(2), 225–237. https://doi.org/10.3758/PBR.16.2.225

Salthouse, T. A., Pink, J. E., & Tucker-Drob, E. M. (2008). Contextual analysis of fluid intelligence. *Intelligence, 36*(5), 464–486. https://doi.org/10.1016/j.intell.2007.10.003

Samani, J., & Pan, S. C. (2021). Interleaved practice enhances memory and problem-solving ability in undergraduate physics. *Npj Science of. Learning, 13*.

Sana, F., Yan, V. X., & Kim, J. A. (2017). Study sequence matters for the inductive learning of cognitive concepts. *Journal of Educational Psychology, 109*(1), 84–98. https://doi.org/10.1037/edu0000119

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*(3), 361–369. https://doi.org/10.1016/j.jarmac.2018.05.005

Schweppe, J., Lenk-Blochowitz, A., Pucher, M., & Ketzer-Nöltge, A. (2024). Interleaved practice in foreign language grammar learning: A field study. *Journal of Educational Psychology*. Online first publication.

Shipstead, Z., Harrison, T., & Engle, R. (2016). Working memory capacity and fluid intelligence: Maintenance and disengagement. *Perspectives on Psychological Science, 11*, 771–799. https://doi.org/10.1177/1745691616650647

Suzuki, Y., Yokosawa, S., & Aline, D. (2022). The role of working memory in blocked and interleaved grammar practice: Proceduralization of L2 syntax. *Language Teaching Research, 26*(4), 671–695. https://doi.org/10.1177/1362168820913985

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

Thurstone, L. L. (1938). Primary mental abilities. *Psychometric Monographs, 1*. ix + 121.

Toppino, T. C., & Bloom, L. C. (2002). The spacing effect, free recall, and two-process theory: A closer look. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 28*(3), 437–444. https://doi.org/10.1037/0278-7393.28.3.437

Unsworth, N. (2016). The many facets of individual differences in working memory capacity. In *, 65. Psychology of learning and motivation* (pp. 1–46). Elsevier. https://doi.org/10.1016/bs.plm.2016.03.001.

Unsworth, N. (2019). Individual differences in long-term memory. *Psychological Bulletin, 145*(1), 79–139. https://doi.org/10.1037/bul0000176

Unsworth, N., Heitz, R. P., Schrock, J. C., & Engle, R. W. (2005). An automated version of the operation span task. *Behavior Research Methods, 37*(3), 498–505. https://doi.org/10.3758/BF03192720

Unsworth, N., & Spillers, G. J. (2010). Working memory capacity: Attention control, secondary memory, or both? A direct test of the dual-component model. *Journal of Memory and Language, 62*(4), 392–406. https://doi.org/10.1016/j.jml.2010.02.001

Wagenmakers, E.-J. (2007). A practical solution to the pervasive problems ofp values. *Psychonomic Bulletin & Review, 14*(5), 779–804. https://doi.org/10.3758/BF03194105

Wahlheim, C. N., Dunlosky, J., & Jacoby, L. L. (2011). Spacing enhances the learning of natural concepts: An investigation of mechanisms, metacognition, and aging. *Memory & Cognition, 39*(5), 750–763. https://doi.org/10.3758/s13421-010-0063-y

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*(2), 285–292. https://doi.org/10.1080/09658211.2019.1705490

Wingert, K. M., & Brewer, G. A. (2018). Methods of studying individual differences in memory. In H. Otani, & B. L. Schwartz (Eds.), *Handbook of research methods in human memory* (1st ed., pp. 443–458). Routledge. https://doi.org/10.4324/9780429439957-24.

Yan, V. X., & Sana, F. (2021). The robustness of the interleaving benefit. *Journal of Applied Research in Memory and Cognition, 10*(4), 589–602. https://doi.org/10.1016/j.jarmac.2021.05.002

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

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