EFFECT OF PRACTICE SCHEDULES ON CONCEPT & CATEGORY LEARNING

THE EFFECT OF PRACTICE SCHEDULES ON THE LEARNING OF CONCEPTS AND CATEGORIES

By FARIA SANA, B.A. (Honours)

A Thesis Submitted to the School of Graduate Studies in Partial Fulfillment of the Requirements for the Degree of Doctor of Philosophy

McMaster University

DOCTOR OF PHILOSOPHY (2015)	McMaster University
Psychology, Neuroscience & Behaviour	Hamilton, Ontario
TITLE:	The Effect of Practice Schedules on the Learning of Concepts and Categories
AUTHOR:	Faria Sana, B.A. (Hons.) (McMaster University)
SUPERVISOR:	Joe Kim, PhD
NUMBER OF PAGES:	xvi; 152

Lay Abstract

Interleaving exemplars from different to-be-learned categories, rather than blocking exemplars by category, often enhances the inductive learning of those categories, as measured by learners' subsequent ability to classify new exemplars from those categories. I examined the generality of the interleaving benefit and the conditions that make interleaving more or less effective for learning than blocking. Consistent with the hypothesis that interleaving enables between-category comparisons, I demonstrate that the interleaving benefit generalizes to the learning of complex, rule-based categories and across all learners, particularly those with lower cognitive abilities. Conversely, blocking enables within-category comparisons, and thus can be as beneficial as, or even more beneficial than, interleaving under certain conditions—if exemplars from the same category are presented at three-at-a-time rather than one-at-a-time or if the categories are structured such that there is high-within and low-between category similarity. These findings highlight the need to shift focus away from examining *which* type of schedule—interleaving or blocking—is more effective for category learning to identifying *when* and *why* each type of schedule may be more effective.

Abstract

Interleaving exemplars from different to-be-learned categories, rather than blocking exemplars by category, often enhances the inductive learning of those categories, as measured by learners' subsequent ability to classify new exemplars from those categories. Majority of the studies on the learning of motor skills, perceptual categories, and mathematics procedures conceptualize the interleaving benefit to be a general learning phenomenon. Results from this dissertation extend the interleaving benefit to the inductive learning of cognitive, rule-based categories (e.g., statistical concepts).

In this dissertation I examine factors that modulate this interleaving benefit, such that interleaving is more or less effective than blocking depending on whether the learning emphasis is on discriminating between categories (*discriminative-contrast hypothesis*) or encoding commonalities within a category (*commonality-abstraction hypothesis*), and depending on whether the temporal spacing between exemplars from the same category optimally promote distributed retrieval practice of critical features shared within a category (*study-phase retrieval hypothesis*). Thus, findings from the current dissertation offer further insight into the boundary conditions of the interleaving benefit.

Consistent with the discriminative-contrast hypothesis, an interleaving benefit was observed when between-category similarity was high and within-category similarity was low, and when there was no temporal spacing between exemplars to disrupt contrast processes critical to between-category comparisons. Consistent with the commonality-abstraction hypothesis, a blocking benefit was observed when between-category similarity was low and within-category similarity was high, and when exemplars were presented three-at-a-time instead of one-at-a-time. Consistent with the study-phase retrieval hypothesis (i.e., introducing spacing between exemplars engages retrieval processes that enhance learning), a blocking benefit was observed when there was temporal spacing between exemplars from the same category.

Moreover, the type of categories themselves and learners' cognitive abilities drove the effects of category learning differently. Findings from the current dissertation begin

to demonstrate the interactions between study schedules and perceptual-based categories (artists' painting styles) and rule-based categories (statistical concepts). For instance, when between-category similarity was low, the interleaving benefit was eliminated for the perceptual-based categories, but no blocking benefit was obtained, contrary to our prediction. This suggests that blocked versus interleaved schedules may be more or less conducive to learning depending on the type of categories. Finally, learners with lower working memory capacities—that is, learners with cognitive limitations related to information processing and integration—benefited from schedules in which exemplars were presented three-at-a-time, and from schedules that were either temporally spaced or interleaved, but having neither or both manipulations produced sub-optimal performance. To conclude, findings from this dissertation clarify when, for whom, and with what kind of categories is interleaving beneficial.

Acknowledgements

I am grateful to the following faculty members and research lab members who have offered their support and feedback during the academic course of my graduate studies:

Members of my supervisory committee: Joe Kim, Scott Watter and Geoff Norman

Members of the department of Psychology, Neuroscience & Behaviour: Karin Humphreys

Members of other psychology departments: Melody Wiseheart, Robert Bjork and Elizabeth Bjork

Members of the Applied Cognition in Education Lab, including Barb Fenesi

Members of the CogFog lab, including Veronica Yan

This research was supported, in part, by the Scholarship from Social Sciences and Humanities Research Council.

Table of Contents

List of Figures

CHAPTER 2

CHAPTER 3

CHAPTER 4

List of Tables

CHAPTER 3

CHAPTER 4

List of Abbreviations

LTM: Long-Term Memory

WM: Working Memory

WMC: Working Memory Capacity

Declaration of Academic Achievement

This sandwich thesis includes three manuscripts, each of which appears in chapters 2, 3 and 4, and all of which have been submitted to scholarly journals for peer-review. Permission from all coauthors to include these manuscripts in my dissertation has been obtained. Once accepted for publication, permission from the respective copyright holders to reprint the manuscripts will be obtained as well.

I conceptualized the theoretical and methodological frameworks, conducted the literature review, collected and analyzed the data, and prepared the manuscripts for submission. The roles of coauthors for each manuscript, the year of data collection, and the conferences where some of the results were presented are documented below.

There is some degree of repetition in the methods sections of manuscripts in chapters 2 and 3, and in the introduction sections of manuscripts in chapters 3 and 4. Each chapter can be read as a stand-alone document, although subsequent chapters do build on previous ones to some degree. The text, tables and figures appear in the form that is submitted for publication in journals.

> **Sana, F.,** Yan, V.X., Kim, J.A. Bjork, E.L., & Bjork, R.A. (submitted). Optimizing the Inductive Learning of Categories: Do the Relative Benefits of Interleaving Versus Blocking Exemplars Vary With a Learner's Working Memory Capacity?

This manuscript, data for which was collected on campus in the 2013-2014 academic year, has been submitted to the journal *Memory & Cognition* and is currently under review. I am the first author. The second and third authors, Veronica Yan and Joe Kim, provided general feedback on the manuscript. The two authors, Robert Bjork and Elizabeth Bjork, offered their theoretical input. Specifically, Robert's expertise in the field of interleaving, and his suggestion to include Vlach's (2010) theoretical account as a possible explanation for the results greatly improved the overall quality of the paper. Previous versions of this manuscript were presented at the 2015 Annual Convention of Psychological Science in New York, and at the 2014 McMaster Symposium on Education & Cognition in Hamilton.

Sana, F., Yan, V.X., & Kim, J.A. (submitted). Study Sequence Matters for the Inductive Learning of Cognitive Concepts.

Data for this manuscript was collected on campus in the 2012-2013 academic year. The manuscript has been submitted to the *Journal of Educational Psychology* for review. I am the first author. The second and third authors, Veronica Yan and Joseph Kim, provided feedback and offered revisions on the manuscript structure and theoretical framework. Previous versions of this manuscript were presented at the 2014 Lake Ontario Visionary Establishment Conference in Niagara Falls, at the 2014 McMaster Symposium on Education & Cognition in Hamilton, and at the 2013 Annual Meeting of Psychonomic Society in Toronto.

Yan, V.X., **Sana, F.,** Kim, J.A. Bjork, E.L., & Bjork, R.A (submitted). Learning Categories from Exemplars: Does the Optimal Schedule of Presentation Vary as a Function of Within-Category Versus Between-Category Similarity of Exemplars?

Data for this manuscript was collected in winter and fall of 2014 at McMaster, UCLA, and Amazon Turk. The manuscript has been submitted to the *Journal of Experimental Psychology: Applied*, where it is currently under review. I am the first coauthor. Veronica Yan played an active role in each phase of the research formulation and manuscript preparation. We originally conducted the experiments covered in this manuscript individually. When we met in January 2014 and discussed our results, we decided to combine our work into one paper. The third author, Joe Kim, provided overall feedback on the manuscript. Elizabeth Bjork offered comprehensive structural revisions. The last author, Robert Bjork, provided general feedback on the manuscript. Previous versions of this manuscript were presented at the 2014 Annual McMaster Education and Cognition Symposium in Hamilton, and at the 2014 Annual Meeting of Psychonomic Society in Long Beach.

Chapter 1: General Introduction

Much of learning—both in childhood and throughout the lifetime—is comprised of learning concepts and categories, which largely involves abstracting general principles from exposure to multiple exemplars of a particular concept or category. Individuals can then apply these principles to categorize new exemplars encountered in the same or different situations. For instance, children are able to abstract or generalize the concept of a chair after seeing a straight-backed chair, a plush armchair, and three-legged stool.

The ability to generalize from exemplars is crucial not only in the natural learning that is required in our day-to-day lives, but also in more formal educational contexts. For instance, illustrating statistical concepts, in a statistics course, through several exemplar problems foster inductive processes that enable students to extract the relevant features that define the concept, and to detect features that differ across other concepts. This type of inductive learning, on a final exam, helps the students to identify (or categorize) a given problem as being of, for instance, an independent *t*-test as opposed to a dependent *t*-test, and then to apply the appropriate solution procedure.

Research on the inductive learning of categories and concepts is important, both theoretically and practically, as it can provide one of the most important windows into the structure of the human mind, and it can allow us to tailor classroom instruction to optimize student learning. The main goals of the current dissertation were to explore instructional methods that optimize the inductive learning of complex categories, to investigate the cognitive mechanisms that make such methods effective, and to examine if the learning advantage conferred by these methods generalizes across domains and across learners of varying cognitive abilities.

The Effects of Practice Schedules on Inductive Learning

The instructional method that I focused on in this dissertation is the relative benefits of interleaved versus blocked study or practice schedules. In a blocked schedule, the learner focuses on learning one skill or concept at a time (e.g., $a_1a_2a_3b_1b_2b_3c_1c_2c_3$, where a, b, and c refer to three different skills or concepts, and 1, 2, and 3 refers to the different exemplars of those skills or concepts). For example, in sports training, the

curler may practice out-turn line of delivery with draw weight several times until some level of competency is achieved, after which he or she may move on to practice the next skill, such as in-turn line of delivery with draw weight, and so on. In educational settings, a student may practice several problems on a single topic until some level of mastery is achieved, after which he or she may move on to practice problems on the next topic, and so on. In an interleaved schedule, different but related skills or concepts are practiced intermixed together (e.g., $a_1b_1c_1b_2c_2a_2c_3b_3a_3$). The curler may practice out-turn and in-turn lines of delivery in a random order, and the student may mix practice problems on current and previously covered topics together.

So which practice schedule is better for the curler and for the student? Intuitively, it makes sense for blocked schedules to produce better skill or concept retention than interleaved schedules. Research on the effects of interleaved and blocked schedules on learning have been studied for the learning of motor skills, category induction, and cognitive procedures, majority of which, contrary to our intuition, report better skill or concept retention under interleaved practice, but some studies also report the opposite—that is, blocked practice produces greater learning gains compared to interleaved practice. These mixed findings indicate that the investigations on when and why interleaved schedules enhance learning are still active and important empirical issues—and, ones that I focused on in this dissertation. In the remainder of this chapter, I provide an overview of the general findings (from various domains) related to practice schedules, the theoretical accounts that explain the interleaving effect, and potential factors that moderate the interleaving benefit, and even show blocking benefits.

Interleaving Benefits Across Several Domains

Findings from a range of research areas including studies on motor skill learning, perceptual category learning, and cognitive procedural learning demonstrate that the interleaving effect may be a general learning phenomenon.

Interleaved Practice Schedules Enhance Psychomotor Skills

Much of the research to date has focused on investigating the benefits of interleaving (or randomizing) practice for the learning of motor skills. In these studies,

participants either practice all the trials corresponding to one variation of a movement before performing another variation (i.e., blocked practice) or they practice all the trials corresponding to several variations of a movement together in a random order (i.e., interleaved practice). Although blocked practice may lead to faster initial gains during acquisition, interleaved practice typically yields superior performance than blocked practice on a final test where participants perform novel variations of the practiced movements. For instance, in a study on baseball batters, Hall, Domingues, and Cavazos (1994) found that random practice of different pitches improved batters' hitting performance more than did blocking practice by pitch (i.e., practice one type of pitch repeatedly until mastered, before moving on to the next). This interleaving benefit has been replicated for other sports (e.g., badminton, Goode & Magill, 1786; volleyball, Jones & French, 2007; baseball, Hall et al., 1994; golf; Brady, 1997), and for keyed timing (Simon & Bjork, 2001, 2002), knot tying (Ollis, Button, & Fairweather, 2005), and musical instrument learning (Abushanab & Bishara, 2013; Stambaugh, 2011).

Interleaved Study Schedules Enhance Perceptual Category Learning

More recent research has examined the relative benefits of interleaved and blocked study schedules for perceptual category induction. In these studies, participants are typically asked to study a series of exemplars from several categories that are presented either randomly or blocked by category, and then asked to classify new exemplars as members of the studied categories on a final classification test. Interleaving exemplars from different categories often produces superior final-test performance than does blocking exemplars by category. For instance, Kornell and Bjork (2008) asked participants to learn artists' painting styles either by presenting different paintings of the same artist in a row (i.e., blocked) or by mixing paintings of different artists (i.e., interleaved) such that no two paintings by the same artist appeared consecutively. Interleaving study produced higher scores than blocking study on a final classification test where participants were shown previously unseen paintings by the studied artists and asked to identify the artist responsible for each new painting. The interleaving benefit has been replicated many times for perceptual category induction not only of artists' painting

styles (Kang & Pashler, 2012; Kornell, Castel, Eich & Bjork, 2010), but also of butterfly species (Birnbaum, Kornell, Bjork &Bjork, 2013) and of bird families (Wahlheim, Dunlosky, & Jacoby, 2011).

Interleaved Practice Schedules Enhance Cognitive Procedural Skills

Research has shown that the benefits of interleaving may also generalize to mathematics learning (e.g., Le Blanc & Simon, 2008; Mayfield & Chase, 2002; Rohrer, Dedrick & Burgess, 2014; Rohrer & Taylor, 2007; Taylor & Rohrer, 2010). In these studies, participants are taught target concepts, including relevant procedures, and then they practice solving word-problems using those procedures. Problems of different procedures are practiced intermixed (i.e., interleaved), or all of the problems of a given procedure are practiced together (i.e., blocked). Following practice, a final problem-solving test is administered in which participants are asked to solve each new word-problem using the appropriate procedure. For instance, Rohrer & Taylor (2007) taught participants to find the volume of four different geometric solids by practicing problems of all four types interleaved with each other or by practicing problems of the same type blocked together. Interleaving practice produced better scores than blocking practice on a problem-solving test where participants were shown new problems and asked to identify the appropriate test type and corresponding formula, and then execute the solution procedure.

Findings from a range of research areas demonstrate that the interleaving effect may be a general learning phenomenon. One goal of this dissertation is to add to this body of literature by demonstrating that the interleaving benefit also extends to the inductive learning of complex, text-based and educationally-relevant categories, namely non-parametric statistical concepts.

Theoretical Accounts of the Interleaving Effect

The two dominant explanations for the interleaving benefit are the study-phase retrieval hypothesis (e.g., Bjork, 1975; Thios & D'Agostino, 1976) and the discriminative contrast hypothesis (Kang & Pashler, 2012; Kornell & Bjork, 2008), each of which highlight the critical components of an interleaved schedule: (1) exemplars from the same

category are spaced, which improves memory and recall of critical features shared within a category, and (2) exemplars from different, but highly similar, categories are interleaved, which enhances induction of critical features that differ across categories. Several studies have been conducted to separate the effects of temporal spacing with those of discrimination learning in order to better understand the mechanisms that contribute to the interleaving effect (Birnbaum et al., 2013; Taylor & Rohrer, 2010). The results seem to suggest that the value of this schedule may come separately from temporal spacing and from discrimination learning (e.g., Birnbaum et al., 2013).

Study-Phase Retrieval Hypothesis

In interleaved study, exemplars from a given category are temporally spaced apart in time (e.g., $a_1...a_2...a_3$), whereas in blocked study, exemplars from a given category are massed together in immediate succession (e.g., $a_1a_2a_3$). One account of the interleaving effect, namely the study-phase retrieval theory, relates the advantage of interleaving to this greater temporal spacing introduced between exemplars from the same category. This account is an extension of the spacing effect, which essentially suggests that repetitions of items further apart in time produce better memory traces of those items than do repetitions close together in time (e.g., Cepeda, Pashler, Vul, Wixted, & Rohrer, 2006).

The study-phase retrieval hypothesis (e.g., Hintzman, 2004; Thios & D'Agostino, 1976), or more broadly, the spacing effect theory, proposes that the interval between exemplars from the same category promotes the forgetting of features included in the exemplars. Because of this forgetting, learners have more difficulty retrieving prior study exemplars, which engages them in greater cognitive effort retrieving the category information, solidifying the memory trace, and slowing the future forgetting rate of that information. Thus, it is the forgetting and subsequent (successful) retrieval of previously studied features that promotes long-term retention of categories. When exemplars are blocked by category, details of previous exemplars for that category are not very old and may still be active in working memory. Therefore, no memory benefits are conferred because participants do not get an opportunity to forget previously studied features and engage in effortful retrieval.

Supporting evidence for this hypothesis comes from a study conducted by Rohrer, Dedrick and Burgess (2014). Unlike most previous studies on the interleaving effect that used highly superficially similar categories—presumably making noticing the differences between them critical for category learning—Rohrer et al. used problems of different types that were superficially dissimilar from each other—making noticing the differences between problem types easily distinguishable to the participants. Participants who practiced problems of multiple types in a given math assignment demonstrated greater learning gains on a delayed test compared to those who practiced all problems of one type at a time. After examining participants' errors, the authors concluded that interleaving practice improved math learning not by decreasing discrimination errors, as would be predicted by the discriminative-contrast hypothesis, but by strengthening the association between each type of problem and its corresponding solution, which presumably occurred because problems of the same type were spaced across assignments (i.e., the spacing effect).

In further support of the spacing effect contributing to the interleaving benefit, Birnbaum et al. (2013) investigated the effect of varying temporal spacing (smaller vs. larger spacing) on interleaved learning while holding juxtapositions constant (i.e., a given category was juxtaposed against the same number of other categories) across all interleaved conditions. Any observed differences in performance could not be due to discriminative-contrast as the degrees of juxtaposition did not vary across conditions, but rather be due to temporal spacing, consistent with the study-phase retrieval hypothesis. Indeed, they found that large spacing produced better classification performance compared to small spacing presumably because greater spacing allowed for more forgetting and subsequently more effortful retrieval.

Discriminative-Contrast Hypothesis

Given that a typical comparison of practice schedules is confounded because interleaving inherently introduces spacing, recent studies have compared the effects of practice schedule on induction while controlling for the effect of spacing by inserting temporal spacing between successive exemplar presentations. Overall, their results

indicate that interleaving has an advantage for learning over and above that of spacing; this advantage, as explained by the discriminative-contrast account, is associated with direct juxtaposition of exemplars from different categories.

The discriminative-contrast hypothesis (e.g., Birnbaum et al., 2013; Kang & Pashler, 2012; Zulkiply & Burt, 2012) proposes that interleaving exemplars from different categories promotes between-category discrimination because it draws learners' attention to the features that vary between categories. When exemplars from one category differ on a number of dimensions from exemplars from another category, juxtaposing their exemplars makes the discriminative features salient. Conversely, blocking exemplars from a category renders it harder to notice salient features differing between categories. Supporting evidence for this hypothesis comes from several studies in the domain of cognitive procedural learning (Taylor & Rohrer, 2010) and perceptual category learning (Birnbaum et al., 2013; Zulkiply & Burt, 2012).

Taylor and Rohrer (2010) compared blocked and interleaved practice of solving four kinds of problems concerning different parts of a prism (i.e., calculating corners, edges, faces, and angles for prisms of varying number of base sides), and found that interleaving produced better performance on a final problem-solving test. Importantly, their data revealed that the observed advantage was likely due to the fact that participants who practiced in a blocked manner made more discrimination errors during the problem solving test—i.e., they used the formula that was appropriate for one of the other kinds of problems. Whereas interleaving provided participants with an opportunity to practice choosing the appropriate formula for a given kind of problem, blocking did not given that every consecutive problem in the blocked schedule concerned the same formula (e.g., Rohrer, 2009; Taylor & Rohrer, 2010).

Interleaving studies on perceptual category learning (Birnbaum et al., 2013, Kang & Pashler, 2012; Zulkiply & Burt, 2012) have manipulated spacing independent of interleaving (and therefore tested the discriminative-contrast hypothesis) by inserting filler tasks between successive exemplar presentations in the blocked conditions and interleaved conditions. In support of the discriminative-contrast hypothesis, they found

that interleaving led to benefits in learning over and above that of temporal spacing—Kang and Pashler (2012) showed that spacing blocked exemplars so that the interval between successive exemplars from the same category was equal to the analogous interval in an interleaved schedule did not lead to the same benefit as interleaved study. Conversely, Birnbaum et al. (2013) and Zulkiply and Burt (2012) found that inserting filler items, to increase temporal spacing, between to-be-studied exemplars decreased the interleaving benefit, presumably because the fillers disrupted the discriminative comparison processes critical to the interleaving effect.

In summary, the theoretical accounts of the interleaving effect suggest two possible mechanisms by which interleaved schedules promote learning: spacing exemplars from the same category enhances memory of critical features shared within a category and juxtaposing exemplars from different categories fosters comparisons that make differences between categories salient. The two accounts may not necessarily be mutually exclusive. In this dissertation, I propose that implementing an interleaved schedule should facilitate the learning of conceptual, rule-based categories (i.e., statistical concepts) through both enhanced discriminative contrast between categories and enhanced memory for retrieved features within a category. More specifically, I examined the notion that interleaving fosters discrimination learning, but temporal spacing can be valuable as well when it does not interfere with the discriminative contrast processes critical to the interleaving effect.

Moderating Factors of the Interleaving Effect

In the above studies, interleaving exemplars from to-be-learned categories, rather than blocking exemplars by category, seemed to enhance learning and category induction. However, findings from early research in the domain of category learning have tended to favor blocked over interleaved schedules of learning. Gagné (1950), for example, found that blocking nonsense form categories led to better performance and fewer errors during the last two trials of acquisition than did interleaving those categories. Note however that this study measured performance (i.e., acquisition), rather than learning, as opposed to the more recent studies. Similarly, Kurtz and Hovland (1956) presented participants with

four categories of geometric patterns, which varied along four relevant dimensions (shape, color, size, and position; a given category was defined by two of these four dimensions) and one irrelevant dimension. Although they found no performance differences between blocked and interleaved study in the classification of exemplars during acquisition, blocking led to better verbalization of the category-defining rules.

Given the mixed findings on the relative benefits of interleaved versus blocked schedules, researchers have shifted their focus from examining *which* type of study schedule—interleaving or blocking—is more effective for category learning to identifying *when* each type of schedule may be more effective for category learning. Findings from these studies offer several possible explanations for the discrepancy observed between blocking versus interleaving benefits. For instance, category similarity structure modulates the interleaving benefit, such that interleaving is more or less effective than blocking depending on whether successful categorization depends more on discriminating between categories (for which interleaving tends to be more effective) or encoding commonalities within a category (for which blocking tends to be more effective; e.g., Carvalho & Goldstone, 2014; Zulkiply & Burt, 2012). The temporal spacing in between successive exemplars during a study schedule has been shown to eliminate the interleaving effect under certain conditions, and thus, is also a moderating factor.

Another potential moderator of the interleaving benefit is working memory capacity of individual learners. Interleaved schedules demand working memory resources. Thus, who benefits from this schedule may depend on how well learners are able to selectively attend to and process relevant incoming information and integrate it with existing information in long-term memory (e.g., Unsworth & Engle, 2007; Unsworth & Spillers, 2010). Finally, the type or nature of categories to be learned—that is, whether rule-based or perceptual-based—could also interact with the interleaving effect. Perhaps the task of learning naturalistic, perceptual-based categories versus the task of learning conceptual, rule-based categories engages different inductive processes (Ashby & Maddox, 2011). In short, it is important to examine when, why, for whom, and with what kind of categories is interleaving beneficial.

Category Similarity Structure Moderates the Interleaving Effect

One factor that appears to affect schedule efficacy is category similarity structure—that is, the similarity relations within and between the categories in a given set of to-be-learned categories. Studies exploring the role of category structure indicate that interleaving (which is assumed to promote between-category comparison) is more effective for low-discriminability categories—that is, where all categories are very similar to one another (e.g., Birnbaum et al., 2013; Kang and Pashler, 2012; Kornell & Bjork, 2008); whereas, blocking (which is assumed to promote within-category comparison) is more effective for cases in which the exemplars from any given category are highly dissimilar from each other and the few features they share are difficult to identify (Kurtz and Hovland, 1956; Whitman and Garner, 1963; Goldstone, 1996; Carpenter and Mueller, 2013). Indeed, this was the pattern of results observed by Carvalho and Goldstone (2014), who manipulated both within- and between-category similarity, and by Zulkiply and Burt (2012), who varied the number of distracting elements to make the defining features more or less difficult to spot.

Carvalho and Goldstone (2014) created blob-shaped categories (defined by a particular notch in one location of the blob) whose exemplars shared few similarities within- and between-categories (low similarity categories) and found that blocking (where categories alternated only 25% of the time) produced better subsequent classification performance than did interleaving (where categories alternated 75% of the time). This pattern of results was reversed for blob-shaped categories whose exemplars shared a high level of similarity with other exemplars in the same category as well as with exemplars in different categories (high-similarity categories). Zulkiply and Burt (2012), who manipulated difficulty of category discriminations, found similar results: namely, that blocking was more effective for learning highly discriminable categories (with presumably low between-category similarity), whereas interleaving was critical for inducing learning of low-discriminability categories (with presumably high between-category similarity).

Together, findings from these studies demonstrate the boundary conditions of the interleaving effect. They suggest that each of the two very different schedules—blocked and interleaved—foster unique category comparisons—juxtaposing exemplars from the same category fosters within-category comparisons and juxtaposing exemplars from different categories fosters between-category comparisons—and one schedule may be more effective than the other depending on the category similarity structures.

The Type or Nature of Categories Moderates the Interleaving Effect

In the category induction literature, the interleaving effect has been shown to generalize across various perceptually rich categories, such as different artists' painting styles (Kornell & Bjork, 2008) that are highly complex visual categories and rely on flexible combinations of brushstroke technique, color palette, and content matter. Compared to such naturalistic categories that may be difficult to define in terms of a rule or even a set of rules, conceptual, text-based categories (e.g., non-parametric statistical tests) could be seen as well defined and more feature- or rule-based. Is it reasonable to expect a similar pattern of results for both perceptual and conceptual types of categories with respect to optimal study schedules?

One reasonable pattern of results to expect would be one based on the assumptions proposed by Ashby and colleagues (Ashby & Maddox, 2005, 2010) who differentiate between "rule-based" category learning (i.e., the learning of categories that can be defined by easily verbalizable rules, considered to be frontally-mediated hypothesis-testing system) and "information-integration" category learning (i.e., the learning of categories that cannot be defined by easily verbalizable rules, considered to be a striatally-mediated procedural-based learning system). They suggest that the optimal study schedules may partly vary depending on whether the categories are perceptual-based (e.g., artists' painting styles) or rule-based (e.g., statistical tests that I used in this dissertation) given that blocking encourages explicit hypothesis-testing (which is not optimal for implicit, information-integration learning) whereas interleaving makes such explicit hypothesis-testing difficult (but which is optimal for learning in cases where the rules are not verbalizable)—blocking is more compatible with rule-based categories

because it promotes the search for rules and hypothesis testing, whereas interleaving makes the use of such a deliberate search strategy difficult, and thus, is more compatible with the implicit learning system thought to underlie the learning of information-integration based categories.

Temporal Spacing Between Successive Exemplars

Blocked and interleaved schedules differ in the amount of temporal spacing between successive exemplars from the same category. Whereas interleaving increases the temporal spacing between exemplars, blocking decreases the temporal spacing between exemplars. Although greater spacing can enhance category induction (e.g., Birnbaum et al., 2013; Vlach et al., 2008), the interleaving effect seems to be contingent upon temporal conditions that do not interfere with between-category discriminations. Birnbaum et al. (2013) inserted 10-sec trivia questions between successive exemplars to compare categorization performance across blocked and interleaved schedules. They found that increasing temporal spacing eliminated the interleaving benefit: adding both interleaving and spacing did not lead to better categorization performance compared to interleaving alone, a pattern, although not significant, also demonstrated by Zulkiply and Burt (2012; i.e., a non-significant benefit of interleaving without spacing versus with spacing). They concluded that an interleaved schedule is effective if the discriminative contrast processes critical to the benefit are not disrupted.

Birnbaum et al. (2013) also reported a spacing effect when the schedule was blocked and exemplars were spaced apart with 10-sec trivia questions. They suggested that spacing in the blocked schedule is valuable for the reasons that have been used to explain spacing effects in noninductive learning—in particular, the study-phase retrieval hypothesis (e.g., Thios & D'Agostino, 1976). That is, delay allows time for forgetting, making retrieval of previous exemplars from memory more difficult, but thereby enhancing learning when such retrievals are successful. In the blocked condition, two exemplars that are spaced through trivia questions belong to the same category and thus, could serve as reminders of each other, whereas in the interleaved condition, the two exemplars do not belong to the same category, and thus, do not share the same category

label and features that could serve as reminders. In another experiment, they also showed that—while keeping the number of juxtaposed categories for any given category consistent—increasing the interval between exemplars from the same category from an average of three intervening trials to an average of 15 intervening trials enhanced category learning, demonstrating again a benefit of spacing (provided juxtapositions were not disturbed).

Other studies that have examined the effect of temporal spacing on learning artists' painting styles failed to demonstrate the spacing effect (Kang & Pashler, 2012; Zulkiply & Burt, 2012). One possible explanation for this discrepancy in findings may be the nature of the categories. Compared to artists' paintings styles that are more "integration-information based" categories, butterfly species used by Birnbaum et al. are more "rule-based" and thus, best supported by blocked schedules. In my dissertation, I used conceptual categories, defined by sets of rules, to examine whether interleaving and blocking are beneficial depending on whether the temporal spacing disrupts the contrast processes critical to the interleaving benefit and whether the temporal spacing causes forgetting but successful retrieval, critical to the blocking (or spacing) benefit.

Individual Differences in Working Memory Capacity

For category or concept learning to occur, attention must be allocated to process-relevant information in Working Memory (WM). If interference, distraction, or processing of incoming information diminishes attention, that information will not be maintained in WM and must be retrieved from long-term memory (LTM). Thus, WM can be considered as consisting of two distinct processes on which learners may differ: (a) controlled attention maintains a few distinct representations for on-line processing in WM, and prevents attentional capture from irrelevant information; and (b) controlled search of LTM retrieves target information (Unsworth & Engle, 2007).

Theoretically, given the involvement of attention as well as both short- and long-term memory in category and concept learning (Lewandowsky, 2011), we would expect WM to contribute to an individual's ability to benefit from interleaved practice. When presented with several exemplars from different categories (e.g., multiple paintings by

each of several artists), some individuals are better able than others to maintain and process task-relevant information (e.g., features of the studied paintings and the differences between them) and to retrieve related information from long-term memory (e.g., recalling a previously studied painting by artist A) in the face of distraction (e.g., paintings by artists B, C, and D). Practically, if we are to recommend the use of interleaving in classroom instruction, we need to know whether interleaving benefits all learners.

One goal of this dissertation was to closely examine the boundary conditions of the interleaving effect, and potential factors that moderate this effect. Specifically, I examined the interactions between study schedules (interleaved and blocked study) and category similarity structure (i.e., the within- and between-category similarities of the set of to-be-learned categories) and between study schedules and temporal spacing in influencing the effectiveness of category learning. I also examined the extent to which interleaved schedules are optimal across two very different types of categories: feature-defined or rule-based textual categories and perceptual-based, less easily verbalizable categories, and across learners with different WMCs.

Overview of Dissertation

In this dissertation, I present three manuscripts in three separate chapters. In chapter 2, I investigated whether the interleaving effect observed in psychomotor learning, perceptual-based induction, and mathematics learning studies extends to the learning of more conceptual, text-based categories, namely non-parametric statistical concepts (i.e., the learning of classifying when each of the three concepts is applicable for analyzing data in different research design scenarios), and whether the interleaving effect similarly benefits learners with varying WMCs. In most statistics courses, students are successful in learning *how* to use procedures for calculating values of certain statistics, such as *t* and *r* values, but they often struggle with learning *when* to use such procedures—that is, the type of problem for which a given statistical procedure is the appropriate analysis. A key aspect of determining when to use a statistical procedure requires building a cognitive structure that represents links and relations between

important statistical concepts. I examined whether an interleaved study schedule would promote a learning sequence that enabled the noticing and learning of structural differences across statistical concepts, particularly when the task was to learn *when* to apply them.

If we are to recommend the use of interleaved practice schedules in classroom instruction, we not only need to know if the benefit generalizes to educationally-relevant materials, but also if this schedule benefits all students similarly. From the standpoint of placing demands on a learner's WMC, interleaving exemplars would appear, intuitively, to be non-optimal, given that noticing the commonalities and differences that define a given category requires remembering the features of prior exemplars from that category and comparing those features with features of other to-be-learned categories. Higher-WMC individuals should be better at maintaining a limited number of representations in the focus of attention and using cue-driven retrieval processes to recover inactive representations from LTM. Such abilities should allow them to compare the to-be-learned features of different categories more effectively and to search LTM more strategically for the features necessary to integrate among the exemplars from a given category. Conversely, lower-WMC individuals may be unable to maintain several features of different categories in WM at once (i.e., they may experience cognitive overload), especially in the presence of contextual interference. Their search of relevant features from LTM may also include irrelevant features, which can impair learning. Issues pertaining to learning across individuals with different working memory capacities are addressed in both chapters 2 and 3.

In chapter 3, I examined the discriminative-contrast and the study-phase retrieval accounts using learners' categorization of conceptual categories and their WMC scores. I tested the discriminative-contrast hypothesis in two ways: by disrupting the contrast processes thought to be critical to the interleaving effect by introducing temporal spacing between successive exemplars, and by making the contrast processes more effective through simultaneous exemplar presentations. If the advantage of interleaving is in part due to discriminative contrast, then disrupting this discrimination process should decrease

classification performance. Also, if the advantage of spacing is due to some forgetting of and subsequent retrieval of critical features, then adding temporal spacing to a blocked schedule should increase classification performance. In further support of the discriminative-contrast hypothesis, interleaving should produce better learning gains under conditions in which learners view exemplars from different categories at once (simultaneously) instead of sequentially. In a sequential schedule, between-category comparisons available to the learner by way of temporal juxtaposition are not as explicit compared to a simultaneous schedule, which may provide a more explicit learning context to elicit the critical differences.

I also examined the commonality-abstraction account, which proposes that when problems of a concept differ on a number of dimensions from other problems of the same concept (i.e., there are within-category differences), juxtaposing or blocking these problems makes the common features shared among the exemplars salient, and thereby facilitates concept induction. Borrowing the framework from research on analogical-reasoning (the process of identifying how aspects of one item correspond with aspects of another item) as a possible theoretical perspective on the role of within-concept comparisons, I argue that making an analogy may be similar to making a within-concept comparison. In other words, determining what can be mapped across two items is a similar process to determining why two problems share the same concept name. Studies in the analogical-reasoning domain have shown that comparing two or more items side-by-side promotes deep processing of the content because their similarities become highlighted, helping learners to abstract principles that may be applied in the future (e.g., Catrambone & Holyoak, 1989; Gentner, 1983; Gick & Holyoak, 1983). Extending these findings now into the category-learning domain, blocking should produce greater learning gains under conditions in which learners view exemplars from the same categories simultaneously rather than sequentially.

With respect to learners' WMC, the study-phase retrieval hypothesis makes strong predictions. Given that the benefit of interleaving depends on successful retrieval or 'reminding' of the prior exemplars, and successful retrieval of relevant features depends

on the learner's ability to engage in controlled and strategic search of LTM, there may be conditions under which retrieval is too difficult (e.g., temporal spacing causes too much forgetting). If to-be-retrieved features of a given category are associated with multiple contexts (e.g., features of other categories), as would be the case during an interleaved schedule with temporal spacing, learners must retrieve target features from LTM through controlled search using only relevant context cues (Unsworth & Engle, 2007), as well as successfully combating interference to prevent intrusions from other contexts—qualities that are observed among individuals with high WMC (e.g., Kane & Engle, 2000). In other words, interleaving benefits may be reduced or eliminated for individuals with low WMC.

In chapter 4, I examined two factors: category similarity structure and the nature of the to-be-learned categories that can potentially moderate the interleaving effect. While in chapter 3, I made the process of making discriminations and comparisons more or less difficult via temporal spacing, in chapter 4, I varied the importance of making between-category discriminations by manipulating category similarity structure and examined the interaction between this structure manipulation and study schedules. Carvalho and Goldstone (2014) manipulated within- and between-category similarity together, and Zulkiply and Burt (2012) focused only on between-category discriminations. Even previous studies that used realistic or naturalistic materials, such as artists' painting styles, consisted only of landscape paintings which were constructed with both high within- and between-category similarity together (Carvalho & Goldstone, 2014, Experiment 1; Kang & Pashler, 2012; Kornell & Bjork, 2008). In order to directly test the hypotheses that a blocked study schedule fosters the noticing of within-category commonalities and an interleaved study schedule fosters the noticing of between-category discriminations, I examined category structures with high within- and low between-category similarity and the inverse, low within- and high-between category similarity, where similarity is defined by irrelevant characteristics of the stimuli (e.g., the surface "cover story" of the research design scenarios/problems or the subjects of the paintings). By making the commonalities or the differences trivially easy to spot (e.g., all of the

research designs appropriately analyzed using a chi-square test concerned cover stories about schooling, or all of the paintings by the artist Grossman consisted of flowers), these hypotheses make very clear predictions for when blocking is beneficial and when interleaving is beneficial.

I also examined whether the optimal study schedules are similar or different depending the nature of the categories that are to be learned. I compared perceptual-based categories (i.e., artists' painting styles) with conceptual-based categories (i.e., statistical concepts). Compared to categories of artistic styles that are perceptually rich, categories of non-parametric statistical tests rely less on perceptual features and more on structural features and rules to abstract commonalities within a category or to distinguish subtle features between categories. Thus, we may see a different pattern of results to suggest that "information-integration" and "rule-based" categories may have different optimal schedules.

Chapter 2: Optimizing the Inductive Learning of Categories: Do the Relative Benefits of Interleaving Versus Blocking Exemplars Vary With a Learner's Working Memory Capacity?

Study Motivation and Overview

In this chapter, I investigated whether the interleaving effect observed in psychomotor learning, perceptual-based induction, and mathematics learning studies extends to the learning of more conceptual, text-based categories, namely non-parametric statistical concepts (i.e., the learning of classifying when each of the three target concepts is applicable for analyzing data in different research design scenarios), and whether the interleaving effect similarly benefits learners with varying WMCs. In most statistics courses, students are successful in learning *how* to use procedures for calculating values of certain statistics, such as *t* and *r* values, but they often struggle with learning *when* to use such procedures—that is, the type of problem for which a given statistical procedure is the appropriate analysis. A key aspect of determining when to use a statistical procedure requires building a cognitive structure that represents links and relations between the important statistical concepts. I examined whether an interleaved study schedule would promote a learning sequence that enabled the noticing and learning of structural differences across statistical concepts, particularly when the task was to learn *when* to apply them.

If we are to recommend the use of interleaved schedules in classroom instruction, we not only need to know if the benefit generalizes to educationally-relevant materials, but also if this schedule benefits all students similarly. From the standpoint of placing demands on a learner's WMC, interleaving exemplars would appear, intuitively, to be non-optimal, given that noticing the commonalities and differences that define a given category requires remembering the features of prior exemplars from that category and comparing those features with features of other to-be-learned categories. Higher-WMC individuals should be better at maintaining a limited number of representations in the focus of attention and using cue-driven retrieval processes to recover inactive representations from LTM. Such abilities should allow them to compare the to-be-

learned features of different categories more effectively and to search LTM more strategically for the features necessary to integrate among the exemplars from a given category. Conversely, lower-WMC individuals may be unable to maintain several features of different categories in WM at once (i.e., experience cognitive overload), especially in the presence of contextual interference. Their search of relevant features from LTM may also include irrelevant features, which can impair learning. In this chapter, I address issues pertaining to learning across individuals with varying working memory capacities.

Optimizing the Inductive Learning of Categories: Do the Relative Benefits of Interleaving Versus Blocking Exemplars Vary With a Learner's Working Memory Capacity?

Faria Sana[1], Veronica X. Yan[2], Joseph A. Kim[1], Elizabeth Ligon Bjork[2], and Robert A. Bjork[2]

[1] *Department of Psychology, Neuroscience & Behaviour, McMaster University*

[2] *Department of Psychology, University of California, Los Angeles*

Author Note

This research was supported by 767-2012-2053 Scholarship from Social Sciences and Humanities Research Council, and by Grant No. 29192G from the McDonnell Foundation. Correspondence concerning this article should be addressed to Faria Sana, Department of Psychology, Neuroscience, & Behaviour, McMaster University, 1280 Main Street West, Hamilton, ON L8S 4K1, Canada. Email: sanaf@mcmaster.ca

Abstract

Interleaving the exemplars of different to-be-learned categories, rather than blocking the exemplars by category, often enhances the inductive learning of those categories, as measured by learners' subsequent ability to classify new exemplars of those categories. From the standpoint of placing demands on a learner's working-memory capacity, however, interleaving exemplars seems non-optimal, given that noticing the commonalities and differences that define a given category requires remembering the features of prior exemplars of that category and comparing those features with features of other to-be-learned categories. In two experiments, we investigated whether individual differences in working-memory capacity moderates the benefit of interleaving. Participants studied exemplars of perceptual categories (artists' styles, Experiment 1) and conceptual categories (non-parametric statistics, Experiment 2), with the studied exemplars presented either blocked by category or interleaved, after which the participants were asked to classify new exemplars of the studied categories. We found a robust benefit of interleaving across both domains and—contrary to our predictions—that benefit appeared for individuals with lower working-memory capacities as well as for individuals with higher working-memory capacities.

Keywords: category induction, learning, interleaving, working memory capacity

Introduction

Contrary to our intuition, research has long demonstrated that practice schedules that randomly sequence (or "interleave") related, but different to-be-learned tasks produce better skill retention than schedules that block practice by task. This finding, referred to as a benefit of contextual interference, or more recently, the interleaving effect, has been well documented across multiple domains and stimuli. Most of the research has focused on investigating the benefits of interleaving (or randomizing) practice for learning motor skills. For instance, interleaving has been shown to be more effective than blocking for learning motor-movement sequences in the laboratory (e.g., Shea & Morgan, 1979; Simon & Bjork, 2001), for complex musical instrument learning (Abushanab & Bishara, 2013), knot-tying skills (Ollis, Button, & Fairweather, 2005) and sports-related abilities (e.g., volleyball, Bortoli, Robazza, Durigon, & Carra, 1992; badminton, Goode & Magill, 1986; baseball, Hall, Domingues, & Cavazos, 1994).

More recently, research has shown that the benefits of interleaving may generalize to learning in the cognitive domain—such as learning when to apply different mathematical formulae (Rohrer, Dedrick, & Burgess, 2014; Taylor & Rohrer, 2010) and inducing natural or perceptual-based categories and concepts (e.g., Birnbaum, Kornell, Bjork, & Bjork, 2013; Kornell & Bjork, 2008). Category induction involves abstracting general principles of a category (or multiple categories) via study or exposure to exemplars. In category-learning studies, participants are typically asked to study a series of exemplars from several categories that are presented either randomly or blocked by category, and then asked to classify new exemplars as members of the studied categories on a final test. Interleaving exemplars from different categories produces better final-test performance than does blocking exemplars by category when learning perceptual categories (e.g., artists' painting styles: Kornell & Bjork, 2008; Kornell et al., 2010; Kang & Pashler, 2012; butterfly species: Birnbaum et al., 2013; bird families: Wahlheim, Dunlosky & Jacoby, 2011), and text-based concepts (e.g., mathematics: Le Blanc & Simon, 2008; Mayfield & Chase, 2002; psychopathological disorders: Zulkiply et al., 2012).

For the learning of some categories/concepts, however, blocking may be more effective than interleaving. Evidence for a blocking benefit comes from tasks that require recognition of within-category commonalities rather than between-category differences (Carvalho & Goldstone, 2014), and when the main challenge is to discover rules that define categories (Noh, Yan, Bjork, & Maddox, under review). Another potential moderator of the interleaving benefit is working memory capacity (WMC) of individual learners, which could influence selective attention and processing of relevant incoming information and its integration with existing information in long-term memory (e.g., Unsworth & Engle, 2007; Unsworth & Spillers, 2010).

Both theoretical and practical reasons exist for examining the relation between individual differences in working memory (WM) and the occurrence of the interleaving effect. Theoretically, given the involvement of attention as well as both short- and long-term memory in category and concept learning (Lewandowsky, 2011), we would expect WM to contribute to an individual's ability to benefit from interleaved practice. When presented with several exemplars of different categories (e.g., multiple paintings by each of several artists), some individuals are better able than others to maintain and process task-relevant information (e.g., features of the studied paintings and the differences between them) and to retrieve related information from long-term memory (e.g., recalling a previously studied painting by artist A) in the face of distraction (e.g., paintings by artists B, C, and D). Another theoretical motivation for examining the relation between WM differences and practice schedules is that several theories of instructional design (e.g., cognitive load theory; for reviews, see Fenesi et al., 2014; Sweller, 1994; Sweller, Merrienboer & Paas, 1998) emphasize the inherent role of WM in learning, but studies directly examining this relation have yet to be conducted.

Practically, if we are to recommend the use of interleaving in classroom instruction, we need to know whether interleaving benefits all learners. In two experiments, we explored the interaction between WM and the interleaving benefit using two very different stimuli: perceptual-based categories (artists' painting styles) to replicate the original Kornell and Bjork (2008) study and conceptual categories (non-

parametric statistics) to extend the work to another type of educationally relevant learning.

We first describe the theories for why interleaving is effective, the general theoretical framework of WM motivating the current research, and then explain how individual differences in WM might contribute to variance in the benefits of interleaved practice.

Why Is Interleaving Better than Blocking?

The two dominant explanations for the interleaving benefit are the discriminative contrast hypothesis (Kang & Pashler, 2012; Kornell & Bjork, 2008) and the study-phase retrieval hypothesis (e.g., Bjork, 1975; Thios & D'Agostino, 1976). The discriminative-contrast hypothesis proposes that interleaving exemplars from different categories promotes between-category discrimination because it draws learners' attention to the features that vary between categories (Goldstone & Steyvers, 2001; Kang & Pashler, 2012). Supporting this theory, Kang and Pashler (2012) found that inserting filler items between to-be-studied exemplars eliminated the interleaving benefit (presumably because exemplars from different categories were no longer juxtaposed).

The study-phase retrieval hypothesis (e.g., Hintzman, 2004; Thios & D'Agostino, 1976), or more broadly, the spacing theory, proposes that the interval between exemplars from the same category promotes forgetting, leading to more effortful retrieval (or, "reminding," Benjamin & Ross, 2011), which then strengthens the learning of that category more than when exemplars are presented consecutively. Critically, the benefit of spacing (or interleaving) rests on successful retrieval of the previous exemplar while studying a subsequent exemplar. Supporting this hypothesis, Birnbaum et al. (2013) found that increasing the interval between successive exemplars from the same category led to better category learning.

Overview of Working Memory

For learning to occur, attention must be allocated to process-relevant information in WM. If interference, distraction, or processing of incoming information diminishes attention, that information will not be maintained in WM and must be retrieved from

long-term memory (LTM). Thus, WM can be considered as consisting of two distinct processes on which individuals may differ: (a) controlled attention, which serves to maintain a few distinct representations (e.g., goal states for the current task, action plans, item representations in list memory tasks) for on-line processing in WM, and to sustain attention and prevent attentional capture from irrelevant information; and (b) controlled search of LTM. The extent to which information can be retrieved from LTM will depend on overall encoding ability, the ability to reinstate the encoding context at retrieval, and the ability to focus search on target information and exclude interfering information (Unsworth & Engle, 2007).

WMC is an individual-difference variable, such that higher-WMC individuals are better than lower-WMC individuals in both maintaining information in WM and in strategically retrieving information from LTM (Engle & Kane, 2004; Unsworth & Spillers, 2010). WMC is typically measured with complex span tasks, where participants immediately recall short lists of items (e.g., letters, words, digits, visuospatial patterns) in serial order, with items interpolated by an unrelated processing task (e.g., verifying equations or judging sentences or symmetry of patterns). Higher scorers, as opposed to lower scorers, are considered to have higher WMC capacities for concurrently processing and storing information at any given time.

Role of Controlled Attention and Maintenance in the Discriminative-Contrast Hypothesis

The discriminative-contrast account posits that interleaving is effective because it promotes discriminative contrast between categories. When exemplars of one category differ on a number of dimensions from exemplars of another category, juxtaposing their exemplars makes the discriminative features salient. Conversely, blocking exemplars of a category renders it harder to notice salient features differing between categories. Learners must exercise considerable attentional control for interleaved study to be effective, so as to maintain relevant features of previously studied exemplars and compare those with exemplars currently being studied (Shea & Morgan, 1979; Shea & Zimny, 1983).

Clearly, interleaved study is cognitively demanding. Given that higher-WMC individuals are better able to control attention for maintenance of relevant information, ignore irrelevant distractions, and withhold habitual responses than are lower-WMC individuals (Unsworth & Engle, 2007), we expect lower-WMC individuals to benefit less from an interleaved schedule. They may be less likely to resolve and benefit from the contextual interference inherent to interleaving, as the number and complexity of features and categories to which they must attend and retain for comparison may exceed their WMC, resulting in impaired inductive learning.

Role of Controlled Search from LTM in Distributed Study-Phase Retrieval Hypothesis

The study-phase retrieval account suggests that during interleaved practice, the intervals between presentations of exemplars from the same category lead to some forgetting of the associated category features, with this forgetting plus the subsequent 'reminding' upon presentation of the next exemplar from that category leading to enhanced category learning. Thus, in contrast to when exemplars are blocked, interleaving them engages learners in more difficult retrieval of prior exemplars, which in turn solidifies the memory trace and slows down the forgetting rate of the retrieved features.

Critically, the benefit of spacing depends on successful retrieval or 'reminding' of the prior exemplars and thus is also dependent on abilities related to controlled search of LTM. Learners must retrieve features of the previous exemplars of a given category from LTM. To the extent that this search process is both effortful and successful, learning is boosted far more than when study is blocked and thus not requiring retrieval from LTM. If this search process is unsuccessful or erroneous, however, we should see impaired category learning.

Attempting to retrieve target information involves controlled and strategic search of LTM, a process that relies on cues to delimit the search set (Unsworth & Engle, 2007). These cues help to discriminate relevant from irrelevant content to reduce the amount of competition at retrieval (Capaldi & Neath, 1995). Lower-WMC individuals may more

often fail to utilize appropriate retrieval strategies to access cues and have difficulty in resolving cue overload (Unsworth, Spillers & Brewer, 2012). Such individuals tend to use noisier context cues that increase their search set size to include both relevant and irrelevant information compared to higher-WMC individuals. If to-be-retrieved features of a given category are associated with multiple contexts (e.g., features of other categories), as would be the case during an interleaved schedule, learners must retrieve target features from LTM through controlled search using only relevant context cues (Unsworth & Engle, 2007), as well as successfully combating interference to prevent intrusions from other contexts—qualities that are observed among individuals with high WMC (e.g., Kane & Engle, 2000). In other words, interleaving benefits may be reduced or eliminated for individuals with low WMC.

Current Study

Theoretically, we might expect interleaving to be beneficial for learners with higher versus lower WMC because higher-WMC individuals should be better at maintaining a limited number of representations in the focus of attention and using cue-driven retrieval processes to recover inactive representations from LTM. Such abilities should allow them to compare the to-be-learned features of different categories more effectively and to search LTM more strategically for the features necessary to integrate among the exemplars of a given category. Conversely, lower-WMC individuals may be unable to maintain several features of different categories in WM at once (i.e., experience cognitive overload), especially in the presence of contextual interference. Their search of relevant features from LTM may also include irrelevant features, which can impair learning.

Both discriminative-contrast and study-phase retrieval accounts predict that interleaving practice should particularly benefit high-WMC learners. A third account, however—the forgetting-as-abstraction account (Vlach, Sandhofer, & Kornell, 2008)—does not make this prediction owing to its assumption that the forgetting induced by the temporal separation of same-category exemplars during interleaved study induces forgetting of irrelevant content details, which in turn promotes abstraction of relevant,

central features. Given the evidence that higher-WMC individuals typically show a reduced forgetting rate, relative to lower-WMC individuals (e.g., Kane & Engle, 2000; Underwood, 1957), the former should be better able to retrieve the specific details of a category, which may not support the 'gist' abstraction necessary for optimal induction. Indeed, Vlach et al. (2008) argued that being able to easily remember specific details of prior exemplars could, in fact, hinder inductive learning.

In the present experiments exploring the relation between individual differences in WMC and the benefits of interleaving, we used perceptual-based categories in Experiment 1 and conceptual (text-based) categories in Experiment 2.

Experiment 1

The goal of Experiment 1 was to determine whether the relative effectiveness of interleaving versus blocking of exemplars during category learning would differ in relation to WMC. The task, taken from Kornell and Bjork's (2008) research, involved learning the styles of 12 different artists from examples of their paintings.

Method

Participants and design. A total of 104 first-year undergraduate students (*Mean age* = 18.96, *SD* = 1.14; 78 females) participated in exchange for course credit, and presentation schedule (blocked vs. interleaved) was randomly manipulated between-subjects.

Participants in each schedule condition were split into low- and high-WMC subgroups (splitting point: median OSPAN score of 45). Data from 12 participants with OSPAN scores equal to or near the median OSPAN value (range: 42–48) were excluded from further analyses in order to differentiate low- and high-WMC groups, leaving 92 participants (*M* age = 18.97, *SD* = 1.87; 74 females). These 92 participants had a mean OSPAN score of 44.40 (*SD* = 17.20, range = 9–75), which did not differ between schedule conditions (blocked = 45.36, interleaved = 43.41; $t(90) = .55$, $p = .59$).

Materials and procedure. The stimuli, taken from Kornell and Bjork (2008), were landscapes or skyscapes by 12 artists: Georges Braque, Henri-Edmond Cross, Judy Hawkins, Philip Juras, Ryan Lewis, YieMei, Marilyn Mylrea, Bruno Pessani, Ron

Schlorff, Georges Seurat, Ciprian Stratulat, and George Wexler. Participants were told that their task was to learn to recognize the styles of 12 different artists, such that they would be able to identify (from a list of names) the artist responsible for new, never-before-studied paintings on the final test.

To eliminate any differential effects that name learning might have on participants with low and high WMC, participants were first familiarized with the artists' names before they began studying the paintings. In the familiarization phase, participants were given 45 s to study the 12 names, and then they engaged in three cycles of two-letter-stem cued recall tests (with immediate feedback). By the third cycle, both low (M = .93, SD = .11) and high (M = .95, SD = .11) WMC participants could recall the 12 artists' names with high accuracy (allowing for spelling errors), $t(91) = 1.18$, $p = .24$.

During the study phase, six paintings by each artist (for a total of 72) were presented sequentially for 3 s each, with the artist's name presented below each painting. In the blocked condition, all paintings by a given artist were presented consecutively with order of the artists and the order of their specific paintings randomized for each participant. In the interleaved condition, the order of all 72 paintings was block randomized with each block of six containing one exemplar from each artist randomly ordered.

After completion of the study phase and a 45-s distraction period of playing Tetris, the final test was administered. Participants were shown four new paintings (all landscapes) by each artist and asked to select, from a list of names, the artist responsible for each painting (for a total of 48 test items). The test images were presented in four blocks, with each block containing one painting of the 12 artists randomly ordered. As soon as participants made their choice by clicking on an artist's name, the next painting was presented. The test was self-paced and included no feedback.

Following the test, the different schedules were described to participants, and they were asked which schedule they thought would be more effective for the learning of the artists' painting styles (interleaved, blocked, or equally effective). They were also asked to rate how difficult it was to generate rules that helped distinguish among the artists (1 =

very difficult; 5 = very easy) and to rate how often they attempted to verbalize distinguishing features for each artist (1 = never, 4 = always).

Finally, participants performed an automated version of the operation span task (OSPAN) (Kane et al., 2004; Unsworth, Heitz, Schrock, & Engle, 2005). Because WMC measures differ widely across studies, we chose a complex span task to assess WM in the present study in line with other studies assessing the relation between WM and higher-order cognition (e.g., Conway et al., 2005). Complex span tasks represent WM as a multi-faceted system that captures variance from different processes subsumed under WM, such as attentional control and controlled search from LTM (Unsworth & Engle, 2007). Briefly, the OSPAN required participants to solve a series of math problems while trying to remember a sequence of unrelated letters, ranging from three to seven letters in length. At the completion of the task, five scores were calculated. The OSPAN score—the measure used herein and the most common one used to index WMC (see Conway et al., 2005)—is the sum of all letters from the letter sets that were recalled completely in the correct order. Full details of task structure and timing can be found in Unsworth et al. (2005).

Results and Discussion

We analyzed correct classification performance with a 2 (schedule: blocked vs. interleaved) x 2 (WMC: low vs. high) between-subjects analysis of variance (ANOVA). Figure 1 shows correct classification percentages for each of the two schedule conditions in relation to individual WMC scores. A significant main effect of schedule was observed, $F(1, 89) = 8.47$, $MSE = .04$, $p = .005$, 95% CI mean difference [0.04, 0.21], $\eta_\rho^2 = .09$, revealing that new paintings by a given artist were more likely to be correctly classified when that artist's paintings had been studied in an interleaved ($M = .52$, $SD = .19$) versus a blocked manner ($M = .39$, $SD = .23$), replicating prior findings. A significant main effect of WMC was also observed, $F(1, 89) = 8.03$, $MSE = .04$, $p = .01$, 95% CI [0.02, 0.18], $\eta_\rho^2 = .08$, indicating that high-WMC participants ($M = .51$, $SD = .23$) correctly classified more new paintings, on average, than did low-WMC participants ($M = .40$, $SD = .19$), a finding consistent with the notion that high- and low-WMC individuals

differ in their ability to control attention and retrieve information from LTM (Unsworth & Engle, 2007). In contrast to initial predictions, however, a schedule X WMC interaction did not occur, $F(1, 89) = .01$, $p = .93$: The interleaved schedule improved classification performance similarly for both high- and low-WMC participants. This pattern thus suggests that interleaving provides general induction benefits across the ability range; the only differences seem due to baseline differences in cognitive abilities. Interleaved study shifted the entire distribution of scores upwards, but did not change the rank ordering of participants.

Finally, regardless of their WMC and classification performance in the two conditions, 62% of the participants judged blocking to be more effective than interleaving, 27% of the participants judged interleaving to be more effective than blocking, and 11% of the participants said that blocking was as good as interleaving. Also, participants with high- and low-WMC did not differ on ratings of how difficult it was to generate rules that helped distinguish among the artists, although high-WMC participants did rate it as numerically easier ($M = 3.07$, $SD = .93$) than did low-WMC participants ($M = 2.81$, $SD = .83$), $t(86) = 1.41$, $p = .16$, $d = .30$. They did significantly differ, however, in how often they reported attempting to verbalize the distinguishing rules, with high-WMC participants ($M = 2.93$, $SD = .91$) making such attempts more often than did low-WMC participants ($M = 2.36$, $SD = 1.09$), $t(86) = 2.62$, $p = .01$, $d = .56$.

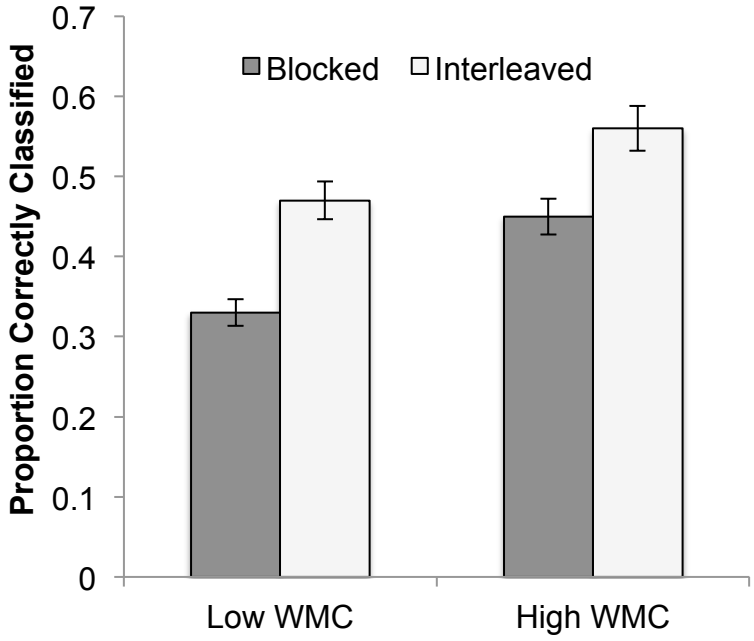

Figure 1. Correct classification proportions for the two presentation schedules and the low- and high-WMC participants in Experiment 1. Error bars represent standard error of the mean.

Experiment 2

As previously indicated, the goals of Experiment 2 were twofold: (a) to see whether the type of scheduling effects observed in Experiment 1 for the induction of artistic styles would extend to the induction of conceptual-based categories—specifically, different statistical tests; and (b) whether any differences in the effectiveness of different study schedules might vary in relation to the WMC of individual learners. In most basic statistical courses, students are successful in learning *how* to use procedures for calculating values of certain statistics, such as *t* and *r* values, but they often struggle to learn when to use such procedures—that is, the type of problem for which a given statistical procedure is the appropriate analysis. Furthermore, this struggle is exacerbated by the fact that most textbooks emphasize how to apply procedures, but not when a procedure is most appropriate (Mayer, Sims, & Tajika, 1995).

A key aspect of determining when to use a statistical procedure requires building a

cognitive structure that represents links and relations between important statistical concepts. One way to help students build such structures is to expose them to different exemplars of concepts that emphasize structural characteristics and require them to categorize these exemplars based on common features. Quilici and Mayer (1996) combined study of worked-out example problems of statistical concepts varying in surface features (e.g., cover stories, objects, numbers) and structural features (relevant for determining the correct solution method) in an attempt to foster students' differentiation between such problem types. Although not manipulated explicitly, students exposed to problem examples in an interleaved fashion, where different methods were used on different problems in the same lesson, outperformed those who were exposed to lessons in which the same method was used on every problem in a lesson. In the present Experiment 2, we extended this work by explicitly examining whether interleaved practice would foster discriminative learning of statistical concepts, particularly when the task was to learn *when* to apply them and, if so, would such an effect vary in relation to individual differences in WMC.

During blocked study, conceptual variability of the word-problems was low, given there was only a single concept on which to focus, and learners needed only to discern the structural features across exemplars within a category and ignore the associated surface features (e.g., cover stories). During interleaved study, conceptual variability was high, with several concepts to consider at once and requiring learners to examine multiple exemplars with varying cover stories of the different concepts presented in a mixed fashion. Thus, in this presentation condition, learners would need to process a greater number of features at any given time, requiring additional WM resources be devoted to deal with this higher cognitive load. Perhaps, then, we would only find interleaving to be a beneficial instructional strategy in this situation when sufficient WMC was available to process the additional features.

Method

Participants and design. A total of 136 first-year undergraduate students (*M* age = 18.63, *SD* = 1.88; 95 females) participated in exchange for course credit. Presentation schedule (blocked vs. interleaved) was randomly manipulated between-subjects.

After splitting the entire group into low- and high-WMC subgroups (splitting point: median OSPAN score of 40), we excluded 11 participants whose scores were equal to or near the median value (range: 38–43) in order to differentiate low- and high-WMC groups. Thus, the final number of subjects in Experiment 2 was 125 (88 females, *M* age = 18.68, *SD* = 1.94). Of these, the mean OSPAN score was 41.24 (*SD* = 16.54, range = 10–75), which did not differ between the schedule-groups (blocked = 42.26, interleaved = 40.22; $t(123) = 1.27$, $p = .204$).

Materials and procedure. The stimuli were three non-parametric statistical concepts: Chi-Squared Test, Wilcoxon Signed-Rank Test, and Kruskal-Wallis Test. Participants studied illustrative word-problems with the instruction to learn to recognize the structural features of the three different concepts, such that they would be able to identify (from a list of names) the concept illustrated in novel word-problems on the final test. Examples of word-problems used in the study and test phases are shown in the Appendix.

During the study phase, all problems (four per concept) were presented sequentially for 25 seconds each with the name of the concept presented above each problem. In the blocked condition, all problems of a given concept were presented consecutively, with their order and the specific problems of a given concept randomized for each individual. In the interleaved condition, the order of all the problems presented for concepts was randomized.

In the final test, which followed a 2-min distractor task (word puzzles), participants were shown three new word-problems of each concept and asked to select, from a list of names, the statistical test/concept that should be used to solve each problem, with the nine test items presented in random order. As soon as participants made their choice on a concept name, the next problem was presented. The test was self-paced and included no

feedback. Once the test phase was complete, participants performed the same automated WM task as used in Experiment 1.

Results and Discussion

We analyzed the final-test classification performance with a 2 (schedule: blocked vs. interleaved) x 2 (WMC: low vs. high) between-subjects ANOVA. Figure 2 shows correct classification proportions for each of the two schedule conditions and the high- and low-WMC groups of participants. As in Experiment 1, a significant main effect of schedule was observed, $F(1, 121) = 22.69$, $MSE = .03$, $p < .001$, 95% CI [0.09, 0.21], $\eta_\rho^2 = .16$, with the interleaved schedule leading to better classification performance than the blocked schedule ($M = .81$, $SD = .16$ vs. $M = .67$, $SD = .21$). The main effect of WMC was also significant, $F(1, 121) = 27.54$, $MSE = .03$, $p < .001$, 95% CI [0.10, 0.22], $\eta_\rho^2 = .19$, with high-WMC participants with correctly classified more test problems than low-WMC ($M = .82$, $SD = .17$ vs. $M = .66$, $SD = .20$).

In contrast to Experiment 1, a significant schedule x WMC interaction was also observed, $F(1, 121) = 6.34$, $MSE = .03$, $p = .013$, $\eta_\rho^2 = .05$. This interaction, however, was in the opposite direction to our initial predictions, with low-WMC participants benefiting more than high-WMC participants from the interleaved schedule. Specifically, low-WMC participants correctly classified more test problems when the schedule was interleaved ($M = .77$, $SD = .18$) than when the schedule was blocked ($M = .55$, $SD = .18$), $F(1, 121) = 24.88$, $p < .001$, $\eta_\rho^2 = .17$; whereas, there was no statistical difference in the performance of the high-WMC-participants in the interleaved ($M = .86$, $SD = .15$) versus the blocked condition ($M = .79$, $SD = .18$), $F(1, 121) = 2.70$, $p = .103$. In fact, while low-WMC participants performed significantly worse than high-WMC participants when study was blocked, $F(1, 121) = 31.84$, $p < .001$, $\eta_\rho^2 = .21$, this difference was almost eliminated when study was interleaved, $F(1, 121) = 3.54$, $p = .062$, $\eta_\rho^2 = .03$.

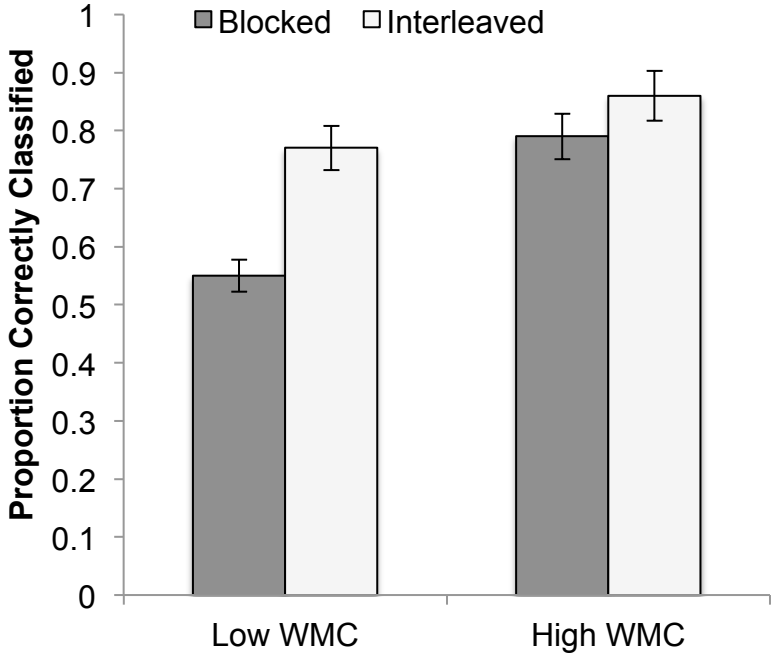

Figure 2. Correct classification proportions for the two presentation schedules and the low- and high-WMC participants in Experiment 2. Error bars represent standard error of the mean.

General Discussion

Across two experiments, we found no evidence that low-WMC individuals benefit less from interleaving than do high-WMC individuals. Rather, we found that interleaving exemplars from multiple categories is significantly more effective than blocking by category for the inductive learning of both types of individuals, and we demonstrated that this effect held in both a perceptual-learning task (artists' painting styles; Experiment 1) and a concept-learning task (statistical tests; Experiment 2). We initially predicted that low-WMC individuals would benefit less from interleaving because interleaved study places greater WM demands on the learner than does blocked study. Instead, however, we found no interaction between schedule and WM in Experiment 1, and a significant interaction in the opposite direction in Experiment 2. Overall, interleaving appeared to be more beneficial for individuals with low WMC than individuals with high WMC; performance difference between the two schedules for low WMC individuals was 22%

and only 7% for high-WMC individuals.

One reason for the different pattern of results between experiments is that performance for high-WMC individuals in Experiment 2 is, to some extent, at ceiling. The pattern of results is consistent, however, with the notion that interleaving homogenizes performance across the ability range—the discrepancy between the high- and low-WMC participants was greatly reduced in the interleaved condition, although not entirely eliminated. One possibility is that high-WMC individuals are already maximizing their cognitive abilities and thus interleaving does not help them much (or at all); low-WMC individuals, however, may benefit from interleaving because they normally do not utilize their cognitive abilities as effectively and interleaving encourages the optimal usage of those processes. Similar findings have been reported with testing (Brewer & Unsworth, 2012; Tse & Pu, 2012) and multimedia instruction (e.g., Sanchez & Wiley, 2006, 2009).

Why did all participants, particularly those with lower WMC, benefit from interleaved schedules when there were compelling theoretical reasons to expect otherwise? The enhanced discriminative contrast associated with interleaving increases the salience of relevant features across categories, a comparison opportunity not afforded in blocked practice, which possibly alleviated limitations to induction associated with having low WMC; that is, lower WMC participants in blocked schedules would have needed to retrieve features of other categories from LTM to make comparisons because exemplars of different categories were not juxtaposed, and the features were likely removed from the focus of attention.

The literature on WM also suggests that individuals with lower WMC tend to use noisier context cues increasing their search set size to include both relevant and irrelevant information, and that they are less successful at combating interference to prevent intrusions from other contexts (e.g., Kane & Engle, 2000). In the current research, the extent to which lower WMC participants attended to, and subsequently retrieved, the relevant features for comparison and contrast in the blocked schedule likely affected induction. A blocked schedule may make it more challenging to notice differences

between categories, whereas an interleaved schedule may make these differences more salient.

Another possible reason why, contrary to our predictions, high-WMC participants did not benefit as much or more from an interleaved schedule than low-WMC participants may be because of their reduced forgetting rates (Vlach et al., 2008). According to the forgetting-as-abstraction hypothesis, the temporal spacing between successive exemplars of a category during an interleaved schedule induces the forgetting of content specific exemplar details. It is this forgetting which then promotes abstraction by reactivating only the central features (rather than specific details) of the prior exemplar on subsequent exemplar presentations of that category. Because high WMC participants show reduced forgetting rates, they may have been better able to retrieve the specific details of an exemplar including those that were not diagnostic of the category, and therefore, not conducive to inductive learning. Conversely, this ability to retrieve specific details of prior exemplars may be especially acute for low-WMC participants, who are frequently susceptible to rapid forgetting. Thus, this group may have been better suited to inductive learning.

The present research adds to the growing body of literature showing the robustness of the interleaving benefit. While knowledge about individual differences in WMC can provide a theoretical basis for tailoring learning and instruction, particularly practice schedules for the learning of categories and concepts, the present studies demonstrate the generalizability of interleaving benefits across very different content domains and across different learner abilities. These findings are important for a number of reasons, not least because they run counter to both learners' intuitions and to prevailing educational practices.

References

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

Benjamin, A.S. & Ross, B.H. (2011). The causes and consequences of reminding. In A. S. Benjamin (Ed.), *Successful remembering and successful forgetting: A Festschrift in honor of Robert A. Bjork* (pp. 71–88). New York, NY: Psychology Press.

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

Bjork, R.A. (1975). Retrieval as a memory modifier. In R. Solso (Ed.), *Information processing and cognition: The Loyola Symposium* (pp. 123–144). Hillsdale, NJ: Lawrence Erlbaum Associates.

Bortoli, L., Robazza, C., Durigon, V., & Carra, C. (1992). Effects of contextual interference on learning technical sports skills. *Perceptual and Motor Skills*, *75*(2), 555–562. http://dx.doi.org/10.2466/pms.1992.75.2.555

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. http://dx.doi.org/10.1016/j.jml.2011.12.009

Capaldi, E.J., & Neath, I. (1995). Remembering and forgetting as context discrimination. *Learning & Memory*, *2*(3-4), 107–132. http://dx.doi.org/10.1101/lm.2.3-4.107

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

Conway, A.R., 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. http://dx.doi.org/10.3758/BF03196772

Engle, R.W., & Kane, M.J. (2004). Executive attention, working memory capacity, and a two-factor theory of cognitive control. *Psychology of Learning and Motivation*, *44*, 145–200. http://dx.doi.org/10.1016/S0079-7421(03)44005-X

Goldstone, R.L., & Steyvers, M. (2001). The sensitization and differentiation of dimensions during category learning. *Journal of Experimental Psychology: General*, *130*(1), 116–139. http://dx.doi.org/10.1037/0096-3445.130.1.116

Goode, S., & Magill, R.A. (1986). Contextual interference effects in learning three badminton serves. *Research Quarterly for Exercise and Sport*, *57*(4), 308–314. http://dx.doi.org/10.1080/02701367.1986.10608091

Hall, K.G., Domingues, D.A., & Cavazos, R. (1994). Contextual interference effects with skilled baseball players. *Perceptual and Motor Skills*, *78*(3), 835–841. http://dx.doi.org/10.2466/pms.1994.78.3.835

Hintzman, D.L. (2004). Judgment of frequency versus recognition confidence: Repetition and recursive reminding. *Memory & Cognition*, *32*(2), 336–350. http://dx.doi.org/10.3758/BF03196863

Kane, M.J., & Engle, R.W. (2000). Working-memory capacity, proactive interference, and divided attention: limits on long-term memory retrieval. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *26*(2), 336–358. http://dx.doi.org/10.1037/0278-7393.26.2.336

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

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

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

http://dx.doi.org/10.1111/j.1467-9280.2008.02127.x

Kornell, N., Castel, A.D., Eich, T.S., & Bjork, R.A. (2010). Spacing as the friend of both memory and induction in young and older adults. *Psychology and Aging*, *25*(2), 498–503. http://dx.doi.org/10.1037/a0017807

Le Blanc, K., & Simon, D. (2008, November). Mixed practice enhances retention and JOL accuracy for mathematical skills. In *49th Annual Meeting of the Psychonomic Society, Chicago, IL*.

Lewandowsky, S. (2011). Working memory capacity and categorization: Individual differences and modeling. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *37*(3), 720–738. http://dx.doi.org/10.1037/a0022639

Mayer, R.E., Sims, V., & Tajika, H. (1995). A comparison of how textbooks teach mathematical problem solving in Japan and the United States. *American Educational Research Journal, 32*(2), 443–460. http://dx.doi.org/10.2307/1163438

Mayfield, K.H., & Chase, P.N. (2002). The effects of cumulative practice on mathematics problem solving. *Journal of Applied Behavior Analysis*, *35*(2), 105–123. http://dx.doi.org/10.1901/jaba.2002.35-105

Noh, S.M., Yan, V.X., Bjork, R.A., & Maddox, W.T. (under review). Optimal sequencing during category learning: Testing a dual-learning systems perspective.

Ollis, S., Button, C., & Fairweather, M. (2005). The influence of professional expertise and task complexity upon the potency of the contextual interference effect. *Acta Psychologica*, *118*(3), 229–244. http://dx.doi.org/10.1016/j.actpsy.2004.08.003

Quilici, J.L., & Mayer, R.E. (1996). Role of examples in how students learn to categorize statistics word problems. *Journal of Educational Psychology*, *88*(1), 144–161. http://dx.doi.org/10.1037/0022-0663.88.1.144

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

Sanchez, C.A., & Wiley, J. (2006). An examination of the seductive details effect in terms of working memory capacity. *Memory & Cognition, 34*, 344–355.

http://dx.doi.org/10.3758/BF03193412

Sanchez, C.A., & Wiley, J. (2009). To scroll or not to scroll: Scrolling, working memory capacity, and comprehending complex texts. *Human Factors: Journal of the Human Factors and Ergonomics Society*, *51*(5), 730–738. http://dx.doi.org/10.1177/0018720809352788

Shea, J.B., & Morgan, R.L. (1979). Contextual interference effects on the acquisition, retention, and transfer of a motor skill. *Journal of Experimental Psychology: Human Learning and Memory, 5,* 179–187. http://dx.doi.org/10.1037/0278-7393.5.2.179

Shea, J.B., & Zimny, S.T. (1983). Context effects in memory and learning movement information. In R. A. Magill (Ed.), *Memory and control of action* (pp. 345–366). Amsterdam: Elsevier. http://dx.doi.org/10.1016/S0166-4115(08)61998-6

Simon, D.A., & Bjork, R.A. (2001). Metacognition in motor learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *27*(4), 907–912. http://dx.doi.org/10.1037/0278-7393.27.4.907

Sweller, J., Van Merrienboer, J.J., & Paas, F.G. (1998). Cognitive architecture and instructional design. *Educational Psychology Review*, *10*(3), 251–296. http://dx.doi.org/10.1023/A:1022193728205

Sweller, J. (1994). Cognitive load theory, learning difficulty, and instructional design. *Learning and Instruction*, *4*(4), 295–312. http://dx.doi.org/10.1016/0959-4752(94)90003-5

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

Thios, S.J., & D'Agostino, P.R. (1976). Effects of repetition as a function of study-phase retrieval. *Journal of Verbal Learning and Verbal Behavior*, *15*(5), 529–536. http://dx.doi.org/10.1016/0022-5371(76)90047-5

Tse, C.S., & Pu, X. (2012). The effectiveness of test-enhanced learning depends on trait test anxiety and working memory capacity. *Journal of Experimental Psychology: Applied, 18*(3), 253–264.

http://dx.doi.org/10.1037/a0029190

Underwood, B.J. (1957). Interference and forgetting. *Psychological Review, 64*(1), 49–60. http://dx.doi.org/10.1037/h0044616

Unsworth, N., & Engle, R.W. (2007). The nature of individual differences in working memory capacity: active maintenance in primary memory and controlled search from secondary memory. *Psychological Review*, *114*(1), 104–132. http://dx.doi.org/10.1037/0033-295X.114.1.104

Unsworth, N., & Spillers, G.J. (2010). Variation in working memory capacity and episodic recall: The contributions of strategic encoding and contextual retrieval. *Psychonomic Bulletin & Review*, *17*(2), 200–205. http://dx.doi.org/10.3758/PBR.17.2.200

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. http://dx.doi.org/10.3758/BF03192720

Unsworth, N., Spillers, G.J., & Brewer, G.A. (2012). Working memory capacity and retrieval limitations from long-term memory: An examination of differences in accessibility. *Quarterly Journal of Experimental Psychology*, *65*(12), 2397–2410. http://dx.doi.org/10.1080/17470218.2012.690438

Vlach, H. A., Sandhofer, C. M., & Kornell, N. (2008). The spacing effect in children's memory and category induction. *Cognition*, *109*(1), 163–167. http://dx.doi.org/10.1016/j.cognition.2008.07.013

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

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

Appendix

Sample of word-problems used in the Study Phase

Wilcoxon Signed-Rank Test

We want to know if the acupuncture treatment is effective to cure chronic back pain. From a random sample, participants are asked to rate their level of back pain before the treatment begins, and again after the treatment ends. Is there a significant change in reported pain after the acupuncture treatment was given?

Chi-Squared Test

We want to know if IPhone users are more likely than Blackberry users to own Apple laptops. From a random sample, participants are categorized as using an iPhone or a Blackberry, and as owning an Apple laptop or a non-Apple laptop. Is using an IPhone related to owning an Apple laptop?

Kruskal-Wallis Test

We want to know if first-born children are happier than their younger siblings. From a random sample, participants are divided into three groups based on if they are the only child, eldest child, or youngest child. All participants report their level of happiness. Is there a difference in happiness ratings among the groups?

Sample of word-problems used in the Test Phase

Wilcoxon Signed-Rank Test

Does drinking alcohol significantly increase the errors in enunciation produced during a singing performance? Professional singers are asked to perform their favourite song before being served with several free tequila shots. They are then asked to perform another song of their choice. Not surprisingly, all singers produced 12 or more enunciation errors once they consumed alcohol.

Chi-Squared Test

Are musicians likely to be more knowledgeable on musical theory than non-musicians? Residents of Ontario, Canada are divided up based on whether they are musicians (i.e., play an instrument for 10+ hours a week) or non-musicians (i.e., do not play any

instrument), and whether they pass or fail an introductory musical theory test. It turns out that musical expertise was not related to musical knowledge.

Kruskal-Wallis Test

Does the choice of the paint colour in a classroom affect student learning? Students are assigned to separate classrooms that have blue, green or yellow coloured paint on the walls and ceilings. They study a chapter on French revolution and then write a multiple-choice test on the content. The test scores did not vary depending on the paint colour of the classroom in which students studied.

Chapter 3: Study Sequence Matters for the Inductive Learning of Cognitive Concepts

Study Motivation and Overview

In this chapter, I examined the discriminative-contrast and the study-phase retrieval accounts using learners' categorization of conceptual categories and their WMC scores. I tested the discriminative-contrast hypothesis in two ways: by disrupting the contrast processes thought to be critical to the interleaving effect by introducing temporal spacing between successive exemplars, and by making the contrast processes more effective through simultaneous exemplar presentations. If the advantage of interleaving is in part due to discriminative contrast, then disrupting this discrimination process should decrease classification performance. Also, if the advantage of spacing is due to some forgetting of and subsequent retrieval of critical features, then adding temporal spacing to a blocked schedule should increase classification performance. In further support of the discriminative-contrast hypothesis, interleaving should produce better learning gains under conditions in which learners view exemplars from different categories at once (simultaneously) instead of sequentially. In a sequential schedule, between-category comparisons available to the learner by way of temporal juxtaposition are not as explicit compared to a simultaneous schedule, which may provide a more explicit learning context to elicit the critical differences.

I also examined the commonality-abstraction account, which proposes that when problems of a concept differ on a number of dimensions from other problems of the same concept (i.e., there are within-category differences), juxtaposing or blocking these problems makes the common features shared among the exemplars salient, and thereby facilitates concept induction. Borrowing the framework from research on analogical-reasoning (the process of identifying how aspects of one item correspond with aspects of another item) as a possible theoretical perspective on the role of within-concept comparisons, I argue that making an analogy may be similar to making a within-concept comparison. In other words, determining what can be mapped across two items is a similar process to determining why two problems share the same concept name. Studies

in the analogical-reasoning domain have shown that comparing two or more items side-by-side promotes deep processing of the content because their similarities become highlighted, helping learners to abstract principles that may be applied in the future (e.g., Catrambone & Holyoak, 1989; Gentner, 1983; Gick & Holyoak, 1983). Extending these findings now into the category-learning domain, blocking should produce greater learning gains under conditions in which learners view exemplars from the same categories simultaneously rather than sequentially.

With respect to learners' WMC, the study-phase retrieval hypothesis makes strong predictions. Given that the benefit of interleaving depends on successful retrieval or 'reminding' of the prior exemplars, and successful retrieval of relevant features depends on the learner's ability to engage in controlled and strategic search of LTM, there may be conditions under which retrieval is too difficult (e.g., temporal spacing causes too much forgetting). If to-be-retrieved features of a given category are associated with multiple contexts (e.g., features of other categories), as would be the case during an interleaved schedule with temporal spacing, learners must retrieve target features from LTM through controlled search using only relevant context cues (Unsworth & Engle, 2007), as well as successfully combating interference to prevent intrusions from other contexts—qualities that are observed among individuals with high WMC (e.g., Kane & Engle, 2000). In other words, interleaving benefits may be reduced or eliminated for individuals with low WMC.

Study Sequence Matters for the Inductive Learning of Cognitive Concepts

Faria Sana[1], Veronica X. Yan[2], and Joseph A. Kim[1]

[1]Department of Psychology, Neuroscience, & Behaviour, McMaster University
[2]Department of Psychology, University of California, Los Angeles

Author Note

This research was supported by 767-2012-2053 Scholarship from Social Sciences and Humanities Research Council. Experiments 2 and 3 of this research study were presented at the annual Psychonomics Conference in 2013.
Correspondence should be addressed to Faria Sana, Department of Psychology, Neuroscience, & Behaviour, McMaster University, 1280 Main Street West, Hamilton, ON L8S 4K1, Canada; Email: sanaf@mcmaster.ca

Abstract

The sequence in which problems of different concepts are studied during instruction impacts concept learning. For example, several problems of a given concept can be studied together (blocking) or several problems of different concepts can be studied together (interleaving). In the current study, we demonstrate that the two sequences impact concept induction differently as they differ in the temporal spacing and the temporal juxtaposition of to-be-learned concept problems, and in the cognitive processes they recruit. Participants studied six problems of three different statistical concepts, and then were tested on their ability to correctly classify new problems on a final test. Interleaving problems of different to-be-learned concepts, rather than blocking problems by concept, enhanced classification performance, replicating the interleaving effect (Experiment 1). Introducing temporal spacing between successive problems decreased classification performance in the interleaved schedule—consistent with the discriminative-contrast hypothesis that interleaving fosters between-concept comparisons—and increased classification performance in the blocked schedule—consistent with the study-phase retrieval hypothesis that temporal spacing causes forgetting and subsequent retrieval enhances memory (Experiment 2). Temporally juxtaposing problems of concepts three-at-a-time rather than one-at-a-time improved overall classification performance, particularly in a blocked schedule—consistent with the commonality-abstraction hypothesis that blocking fosters within-concept comparisons (Experiment 3). All participants also completed a working memory capacity (WMC) task, findings of which demonstrate that the efficacy of the above study sequences may be related to individual differences in WMC.

Keywords: induction, categorization, interleaving, math learning

Introduction

When students are introduced to a concept, the instruction is often paired with several illustrative problems. Exposure to these problems enables them to abstract general principles that define the concept, and to subsequently apply the abstracted principles to novel situations. For instance, the graduate student who has designed several experiments should be able to recognize when a novel research problem requires the application of an independent *t*-test, and be able to differentiate that problem from problems that require the application of other statistical tests.

Importantly, the sequence in which the problems are presented during instruction affects concept learning. For example, in a blocked schedule, several problems within a single concept (e.g., the independent *t*-test) are studied consecutively to extract the key features before moving onto the next concept. This can be contrasted with an interleaved schedule, in which several problems of different concepts are studied together (e.g., independent *t*-test, dependent *t*-test, ANOVA) to learn the subtle differences that exist among them. Each study schedule method contributes to learning differently. While majority of the research suggests that interleaved schedules produce greater learning gains than blocked schedules (see Rohrer, 2012 for a review), some research has shown that blocked schedules also have the potential to optimize concept learning. Understanding the conditions when each of the two schedules is effective may provide a theoretical basis for tailoring learning and instruction.

In the current study, we examined the effects of blocked and interleaved schedules on the learning of statistical concepts. Specifically, we examined the different factors and processes that determine *when* (temporal factors) and *how* (cognitive processes) one schedule may be more or less effective than the other.

Prior Relevant Research

The effects of interleaved and blocked schedules have been studied for the learning of motor skills, mathematics procedures, and for perceptual and text-based categorization. Interleaving is often more effective than blocking, but blocking can also be effective under certain conditions (see Carvalho & Goldstone, 2015 and Rohrer, 2012

for reviews). There are several factors that determine when one schedule is more effective than the other. In the current study, we focused on two temporal factors—juxtaposition and spacing—that may interact with study schedules to influence concept learning. In what follows, we summarize research on the generality of the interleaving benefit, and then review evidence that suggests that the temporal juxtaposition of concept problems and the temporal spacing between concept problems promote unique cognitive processes that differentially drive interleaving and blocking benefits.

Generality of the interleaving effect. Interleaved practice has been shown to promote better procedural learning than blocked practice, in particular for learning mathematics procedures (e.g., Le Blanc & Simon, 2008; Mayfield & Chase, 2002; Rohrer & Taylor, 2007; Taylor & Rohrer, 2010). For instance, in a mathematics learning study, participants learned to find the volume of four different geometric solids by practicing problems of all four types in an interleaved fashion or by practicing problems of the same type together in a blocked fashion (Rohrer & Taylor, 2007). Interleaving practice produced better scores than blocking practice on a problem-solving test where participants were shown new problems and asked to identify the appropriate test type, recall the corresponding formula, and then execute the solution procedure.

Similarly, in a perceptual category induction study, Kornell and Bjork (2008) asked participants to learn artists' painting styles either by presenting different paintings of the same artist in a row (i.e., blocked) or by mixing paintings of different artists (i.e., interleaved) such that no two paintings by the same artist appeared consecutively. Interleaving study produced higher scores than blocking study on a final classification test where participants were shown previously unseen paintings by the studied artists and asked to identify the artist responsible for each new painting. This result has been replicated many times for perceptual category induction not only of artists' painting styles (Kornell, Castel, Eich & Bjork, 2010; Kang & Pashler, 2012), but also of butterfly species (Birnbaum, Kornell, Bjork & Bjork, 2013) and of bird families (Wahlheim, Dunlosky, & Jacoby, 2011). The range of research areas studied suggests that the interleaving effect is a general learning phenomenon. One goal of the current paper was to add to this body of

literature demonstrating that the interleaving effect also generalizes to the inductive learning of statistical concepts.

Role of Temporal Juxtapositions in Fostering Unique Concept Comparisons

Blocking and interleaving differ in the temporal juxtapositions of problems of to-be-learned concepts. In an interleaved schedule, problems of different concepts are juxtaposed together, which fosters between-concept comparisons, consistent with the discriminative-contrast hypothesis. In a blocked schedule, problems of the same concept are juxtaposed together, which fosters within-concept comparisons, consistent with the commonality-abstraction hypothesis.

Interleaving enables between-category comparison: the discriminative-contrast hypothesis. The discriminative-contrast account proposes that when problems of a concept differ on a number of dimensions from problems of another concept, juxtaposing these problems through interleaving makes the discriminative features salient, which facilitates concept induction (Birnbaum et al., 2013; Kornell & Bjork, 2008). Conversely, blocking problems of a concept together renders it harder to notice salient features that differ between concepts, which can make this schedule less effective under these conditions.

Evidence for the discriminative-contrast hypothesis comes from a study by Taylor and Rohrer (2010) in which participants learned to calculate the volume of four different types related to a prism (i.e., corners, edges, faces, and angles) by practicing problems of all four types in an interleaved fashion or by practicing problems of the same type together in a blocked fashion. Interleaving practice produced better scores than blocking practice on a problem-solving test where participants were shown new problems and asked to identify the appropriate problem type, recall the corresponding formula, and then execute the solution procedure. The authors concluded that the observed advantage was due to the fact that participants who practiced in a blocked manner made more discrimination errors during problem solving (i.e., they used the formula that was appropriate for one of the other kinds of problems). Thus, interleaving provided participants with an opportunity to practice how to execute a solution procedure, *and*

discriminate when a given formula was appropriate, whereas blocking did not provide the same opportunity given that every consecutive problem in this schedule concerned the same formula (e.g., Rohrer, 2009; Taylor & Rohrer, 2010).

Perhaps the strongest evidence for the discriminative-contrast hypothesis comes from a study by Birnbaum et al. (2013, learning of butterflies' species). They attempted to interrupt the contrast processes presumed to be critical to the interleaving benefit by inserting unrelated fillers between exemplars. Indeed, performance decreased in the temporally spaced interleaved schedule compared to the typical uninterrupted interleaved schedule. This finding provides evidence for the role of category comparisons in the benefit of interleaving given that making comparisons is more difficult in interleaved schedules with temporal spacing.

Blocking enables within-concept comparison: the commonality-abstraction hypothesis. Commonality-abstraction account proposes that when problems within a concept differ from each other on a number of dimensions, juxtaposing these problems through blocking makes the common features salient, which facilitates concept induction. Conversely, interleaving problems of a concept with problems of other concepts renders it harder to notice salient features that are common within a concept, which can make this schedule less effective under these conditions.

Evidence for the commonality-abstraction hypothesis comes from a series of recent studies that manipulated concept discriminability. Their results suggest that blocking is particularly effective for high-discriminability concepts where all the problems of the same concept are highly dissimilar and the few features they share are difficult to identify (Kurtz and Hovland, 1956; Whitman and Garner, 1963; Goldstone, 1996; Carpenter and Mueller, 2013). For instance, Carvalho and Goldstone (2014) created artificial categories for which the exemplars shared very few similarities within and between categories. They found that blocking the exemplars of a category produced better subsequent classification performance than did interleaving exemplars of all categories together. Similarly, Zulkiply and Burt (2012), who manipulated difficulty of category discriminations, also found that a blocked schedule was more effective for

learning highly discriminable categories (with presumably low between-category similarity).

The findings discussed so far are mainly from the research on category and concept learning, which report limited benefits of blocking. Research on analogical-reasoning, the process of identifying how aspects of one item correspond with aspects of another item, can also provide a theoretical perceptive on the role of within-concept comparisons. Making an analogy may be similar to making a within-concept comparison—determining what can be mapped across two items is similar to determining why two problems share the same concept name. Studies in this domain have shown that comparing two or more items promote deep processing of the content because their similarities become highlighted, helping learners to abstract principles that may be applied in the future (e.g., Catrambone & Holyoak, 1989; Gentner, 1983; Gick & Holyoak, 1983). In their study, Gentner, Loewenstein, and Thompson (2003) asked participants to either compare two cases of a negotiation principle or study each case independently. Those who engaged in the comparison developed better representations of the principle, and were better able to identify and solve novel cases that required the same negotiation principle. Gentner et al. argued that comparisons allowed participants to discover the underlying structure shared by both cases.

The temporal juxtaposition of problems of to-be-learned concepts fosters unique comparisons across problems of different concepts in the interleaved schedules and across problems of the same concept in the blocked schedules. One goal of the current study is to further test the discriminative-contrast and commonality-abstraction hypotheses. In most of the studies, problems are presented one at a one, and the concept comparison available to the participants by way of temporal juxtaposition is not explicitly invited. Studies in the analogical-reasoning domain highlight the importance of encouraging learners to explicitly compare items during instruction. In fact, simply having two cases be presented side-by-side rather than on separate pages fostered the abstraction of the underlying structure (e.g., Gentner et al., 2003; Kurtz, Miao, & Gentner, 2001). We predict that both interleaving and blocking should produce greater learning gains under

conditions in which participants are allowed to simultaneously, rather than sequentially, view problems of concepts, as simultaneous sequences provide a more explicit learning context to elicit the critical differences between concepts and commonalities within a concept.

Role of Temporal Spacing in Enhancing Memory of Concept Features

Blocking and interleaving differ in the amount of temporal spacing that exists between problems of to-be-learned concepts. In an interleaved schedule, problems of the same concept are temporally spaced apart with problems of other concepts interposed between them. In a blocked schedule, problems of the same concept are presented consecutively with no temporal lag or spacing between them. It is this temporal spacing in the interleaved schedules that produces greater learning gains, as proposed by the study-phase retrieval account (Thios & D'Agostino, 1976; Bjork, 1975). This account, an extension of the spacing effect (Cepeda et al., 2006; Dempster, 1988), proposes that the interval (or temporal spacing) between problems of the same concept promotes forgetting, leading to more effortful retrieval, which then strengthens the learning of that concept more than when problems are presented consecutively, such as in the blocked schedules. Critically, the benefit of spacing (or interleaving) rests on successful retrieval of the previous problem while studying subsequent problems of the same concept.

Evidence for the role of temporal spacing in the interleaving benefit comes from a single perceptual category induction study. Birnbaum et al. (2013; the inductive learning of butterfly species) investigated the effect of varying temporal spacing (smaller vs. larger spacing) on interleaved learning while holding juxtapositions constant (i.e., a given category was juxtaposed against the same number of other categories) across all interleaved conditions. Any observed differences in performance could not be due to discriminative-contrast as the degrees of juxtaposition did not vary across conditions, but rather be due to temporal spacing, consistent with the study-phase retrieval hypothesis. Indeed, they found that large spacing produced better classification performance compared to small spacing.

Birnbaum et al. (2013) also reported a blocking benefit when exemplars in the blocked schedule were spaced apart with 10-sec trivia questions compared to when there was no spacing. They suggested that the temporal delay allowed time for forgetting, making retrieval of previous exemplars from memory more effortful, but thereby enhancing learning when such retrievals were successful.

There are several other studies that demonstrate a benefit of temporal spacing, often referred to as the spacing effect (e.g., Cepeda et al., 2006, 2008, 2009)—the robust finding that repetitions of items further apart in time (i.e., spaced) produce better memory than do repetitions close together in time (i.e., massed). However, these studies are different from the studies that examine the benefits of interleaved versus blocked schedules. In a typical interleaving study, problems of other concepts are present (i.e., interleaved) between problems of the same concept. In a typical spacing study, two items of the same category are spaced apart with no other items from different categories between them. Essentially, the blocked schedule with temporal spacing in the current context (e.g., Birnbaum et al., 2013) is essentially the same as a typical spaced condition in a spacing effect experiment. Contrary to the blocking benefit reported by Birnbaum et al., others that have investigated blocked schedules with and without temporal spacing, however, did not find a spacing benefit when learning artists' painting styles (Kang & Pashler, 2012; Zulkiply & Burt, 2012).

Based on the findings, it seems that the spacing inherent in an interleaved design is thought to contribute to the interleaving benefit. Indeed, increasing spacing in interleaved schedules, if it does not disrupt discriminative-contrast processes, actually enhances category induction. Blocked schedules may also benefit if spacing is added, however, evidence for this assertion is rather mixed. Thus, one goal of the current study is to examine the effect of increasing temporal spacing in blocked and interleaved schedules. Increasing temporal spacing in the interleaved schedule should harm learning, as the spacing would interrupt between-concept comparisons critical to the interleaving effect. On the other hand, increasing temporal spacing in the blocked schedule should enable a blocking benefit, consistent with the study-phase retrieval hypothesis. The

spacing between problems of the same concept would allow for some forgetting of and subsequent retrieval of critical features, strengthening the memory of the concept's features.

Overview of the Present Experiments

The ability to recognize when a given statistical concept applies to a research problem involves having knowledge of the relations between the problem structure and the concept features. In the present study, we examined the different study sequence conditions under which this type of concept knowledge is acquired.

In Experiment 1, we examined whether the interleaving effect found with the learning of mathematics concepts and of perceptual categories also extends to the learning of statistical concepts. In the remaining experiments, we focused on two temporal factors—spacing and juxtaposition—that provide insight into the cognitive mechanisms—comparisons and spaced retrieval—that make the study sequences more or less effective.

In Experiment 2, we examined the interaction between study schedules (blocked versus interleaved schedules) and temporal spacing (spacing versus no spacing) to test the comparison and study-phase retrieval hypotheses. Increasing temporal spacing in the interleaved schedule should harm classification as the spacing interrupts between-concept comparisons critical to the interleaving effect (evidence for discriminative-contrast). Conversely, increasing temporal spacing in the blocked schedule should enhance classification, as the spacing allows for some forgetting of and subsequent retrieval of critical features, strengthening the memory of the concept's features (evidence for study-phase retrieval).

In Experiment 3, we examined the interaction between study schedules (blocked versus interleaved schedules) and temporal juxtaposition (sequential versus simultaneous sequences) to test the two comparison hypotheses. Presenting problems simultaneously rather than sequentially should enhance between-concept comparisons in the interleaved schedule (evidence for discriminative-contrast) and within-concept comparison in the blocked schedule (evidence for commonality-abstraction).

In all experiments, we were concerned with the inductive learning of critical features that define the to-be-learned statistical concepts. Participants did not receive any explicit instruction or lessons on the to-be-studied concepts; learning was essentially discovery-based, and the focus was on successfully classifying never-before-seen problems based on the concept representations acquired from studying the problems and inducing the critical features. We chose these training materials because the appropriate application of each of these concepts relies upon a conjunction of features, some of which may overlap with the features of the other concepts. It represents, therefore, a clear case in which learners must first discover the critical features by comparing study problems, and in which memorizing the conjunction of features is not trivial.

A secondary goal of the current study was to examine the relation between participants' cognitive abilities and the sequence conditions to determine if certain study sequences mediate individual differences in cognitive abilities. Working memory (WM) is a stable cognitive trait and a well-established predictor of academic learning (Alloway & Alloway, 2010). Thus, we also included a single WM task to examine whether individual differences in Working Memory Capacity (WMC) predict classification performance in blocked and interleaved study schedules. WMC reflects an individual's ability to actively maintain and process task-relevant information and retrieve related information from long-term memory (LTM) in the face of distraction (Baddeley & Hitch, 1974; Engle & Kane, 2004; Unsworth & Engle, 2007; Unsworth & Spillers, 2010).

Individual differences in WMC can provide theoretical insights into the cognitive processes involved in study sequences, particularly with respect to spaced retrieval (as proposed by the study-phase retrieval hypothesis). For instance, if the inductive learning of statistical concepts depends on successful retrieval of previous problem features, then an individual's ability to strategically search for those features in LTM will play an important role—those with higher WMC may be better able to successfully retrieve previously studied problem features from LTM than those with lower WMC. Furthermore, to the extent that our manipulations vary the memory load required to make comparisons between study problems, individual differences in WMC may play an

important role here as well—those with higher WMC may be better able than those with lower WMC to actively maintain relevant features of a prior problem and, at the same time, attend to the current problem for comparisons. Practically, as a stable characteristic, any differences between higher and lower WMC participants would have direct implications for tailoring instruction in classrooms.

Experiment 1

The goal of Experiment 1 was to examine if an interleaved study schedule promotes the inductive, classification learning of statistical concepts. Participants were presented with several study problems of three statistical concepts that were blocked by concept or interleaved with problems of other concepts. With no lessons or descriptions of these concepts provided beforehand, they had to study the problems closely in order to extract the critical features diagnostic of the concepts. On a final classification test, participants identified the statistical concept that could best be applied to never-before-studied test problems. They all completed the ospan task after the final test.

We predicted that interleaving would produce better classification performance than blocking for two reasons: First, juxtaposing problems of different concepts in the interleaved schedule would promote comparison between the concepts, as features abstracted from the immediately prior problem would still be active in WM and likely available for further processing and integration (discriminative-contrast hypothesis). Second, the temporal spacing between problems of the same concept in the interleaved schedule would promote forgetting of related features, leading to more effortful retrieval of these features from LTM the next time another problem of the same concept is presented. This forgetting and subsequent retrieval across problem presentations would strengthen the learning of that concept and its features (study-phase retrieval hypothesis).

We also examined the relation between participants' WMC and classification performance in each study schedule. Participants with higher WMC would likely do well regardless of the study schedule, as these individuals are generally better at allocating their attention to relevant features and resisting interference during encoding, and at

engaging in controlled LTM search for relevant features that are no longer active in WM during retrieval (Unsworth & Engle, 2007).

We expected participants' WMC to predict their classification performance, particularly in the blocked schedule. The extent to which individuals successfully retrieve encoded information from LTM depends on their ability to engage in controlled and strategic cue-dependent search of LTM (Unsworth & Engle, 2007). Given that low-WMC individuals encode contextual cues at a global level rather than a specific level, their search set is often noisier, which increases intrusions and errors (Kane & Engle, 2000; Rosen & Engle, 1998). Based on these findings, perhaps low-WMC participants may benefit more from an interleaved schedule as it helps to constrain their search set to include more diagnostic features that repeatedly occur across subsequent problems of a concept rather than the irrelevant non-overlapping features.

Method

Participants. One hundred thirty six undergraduate students (95 females; *M* age = 18.62 years, *SD* = 1.88) from McMaster University participated in the experiment in exchange for 1 course credit. There were 68 participants in each of the two conditions (blocked and interleaved). There were no significant differences in age or statistical background between the two conditions. Four additional participants completed the experiment, but were excluded from the analysis for indicating previous knowledge of the statistical concepts being tested in the experiment.

Table 1

Structural features of three statistical concepts used in all experiments.

Statistics test	**Structural features**			
	IV (# of groups)	Sample	DV	Main question
Kruskal-wallis test	3+	Independent	Quantitative	Are there any differences between three or more conditions?
Wilcoxon signed-	1	Dependent	Quantitative	Is there a significant change in

rank test				a condition after some treatment?
Chi-squared test	2	Independent	Categorical	Is there a relationship between two variables?

Materials. All materials were modified from an undergraduate statistics textbook (Gravetter & Wallnau, 2008). Participants learned three non-parametric statistical concepts: Chi-squared test, Kruskal-wallis test, and Wilcoxon signed-rank test. Each of the three concepts was illustrated with six different study problems with research design descriptions for each one of the three statistical tests would be appropriate), none of which included worked-out solutions (see Appendix A). Table 1 lists the four defining features for each concept embedded in the study problems. The content of these problems was similar in structure, length and difficulty. Interleaving has been shown to be most advantageous for categories with low discriminability (Zulkiply & Burt, 2012); materials from the current studies can be classified (although not objectively) as having low to moderate discriminability as the defining dimensions of each statistical concept have multiple possible features with no one feature that can discern the concepts.

In addition to the 18 study problems, there were a total of nine test problems (three for each concept) on the classification test (see Appendix B). Participants had to identify which of the studied concepts best represented a given test problem. These problems were more complex than the ones in the study phase in two ways: there was no one defining feature that cued participants to the correct response, and they had to have learned at least two of the four diagnostic features of a concept in order to get the correct response. Each test problem included at least one distractor feature of a non-target concept.

All study and test problems had different storylines. To ensure that each problem represented only one of the three concepts and included all features defining the target concept, two raters (PhD candidates from the Department of Statistics, McMaster University) independently classified the problems based on the concept they illustrate.

We used Cronbach's α (r = .76) to calculate the inter-rater agreement, and made revisions to the problems with low item agreement.

We used the Automated Operation Span task (ospan; Unsworth et al., 2005) to measure participants' working memory capacity (WMC)—their ability to simultaneously process and store information. During the task, a mathematical operation was presented on the screen (e.g., "Does (4/2) + 1 = 6?"), and participants pressed a key to indicate whether the equation was correct or incorrect. All responses and response latencies were recorded. Following this answer selection, a letter was presented for 800 msec (e.g., 'F'). After a series of two to six operation-letter pairs, participants were asked to recall the list of two to six letters in the order they were presented. Each participant was presented with three sets of each length. A response was counted as correct only if the letter was in the correct serial order, and the letter itself was correctly recalled. The ospan score was the sum of recalled letters for all sets recalled correctly (completely and in order of presentation) with possible scores ranging from 3–75. Full details of the task structure, timing and scoring can be found in Unsworth et al. (2005).

Design and procedure. In this one-hour experiment, participants, tested individually, completed the study phase and test phase, separated by a 2-min distractor task, and then completed the ospan task. In the study phase, participants were to learn three basic concepts, each of which would be illustrated with a series of short word-problems. They were instructed to carefully study these problems and try to identify the features diagnostic of the concepts. Participants were not explicitly told what dimensions defined a concept leaving them to discover on their own what features of the problems were most relevant. In the test phase, they would be presented with a series of new word-problems and their task would be to identify the concept that can best be applied to solve each problem.

We manipulated study schedule (blocked vs. interleaved) between-subjects. In the blocked condition, study problems were blocked by concept (e.g., AAAAAABBBBBBCCCCCC), such that a number of problems for a given concept would appear consecutively. In the interleaved condition, study problems of different

concepts were interleaved (e.g., ABCABCABCABCABCABC), such that no two problems of a given concept appeared consecutively. For both conditions, each problem was presented for 20-sec, one at a time, with a 1-sec blank screen in between presentations. The name of the corresponding statistical concept appeared directly above each problem. After the study phase, participants played a game (minesweeper) for two minutes followed by a self-paced classification test. Test problems, randomly ordered, were presented one at a time with three-alternative forced choice options of the target concepts. Participants selected the concept that best applied to a problem. No feedback was provided.

After completing the classification test, they played a video game (bejeweled) for two minutes, followed by instructions on completing the ospan task. They were told that this task required participants to solve a series of math problems while trying to remember a sequence of unrelated letters, ranging from three to seven letters in length. They also received detailed instructions on screen and practiced the task with feedback before it began. Once they completed the ospan task, participants were debriefed and dismissed.

Results and Discussion

Classification performance. We examined the effect of schedule (blocked vs. interleaved) on classification performance using a between-subject analysis of variance (ANOVA). The results are illustrated in Figure 1. Consistent with the prior studies, the interleaved condition (M = .81, SD = .18) yielded significantly better classification performance than the blocked condition (M = .66, SD = .21), $F(1,134) = 20.29$, MSE = .04, $p < .001$, $\eta\rho^2 = .13$.

Individual differences. We also examined whether differences in classification performance between blocked and interleaved conditions varied as a function of participants' WMC. Figure 2 shows the linear regression lines of the two study schedules as they relate to classification performance and WMC. Classification performance on the final test increases as participants' WM scores increase, with a steeper slope in the blocked condition than in the interleaved condition.

The linear regression analyses demonstrate that participants' WMC predicted their classification performance in the blocked condition, $t(1, 66) = 5.04$, $p < .001$, $R^2 = .28$. This schedule produced non-optimal classification performance, particularly for lower WMC participants, for reasons that are proposed by the study-phase retrieval hypothesis—the first problem of the concept was presumably not retrieved at the time of the second or subsequent problems as the memory traces for all problems were still active in WM.

Participants' WMC only marginally predicted their classification performance in the interleaved condition, $t(1, 66) = 2.01$, $p = .048$, $R^2 = .06$, suggesting that differences in cognitive ability may be mitigated with an interleaved schedule—lower WMC participants profited more than higher WMC participants from studying problems of different concepts together. The overall finding that lower WMC participants primarily drove these results is not surprising, as the evidence suggests that individuals with higher WMC already use efficient strategies to process information unlike their lower WMC counterparts (e.g., Brewer & Unsworth, 2012).

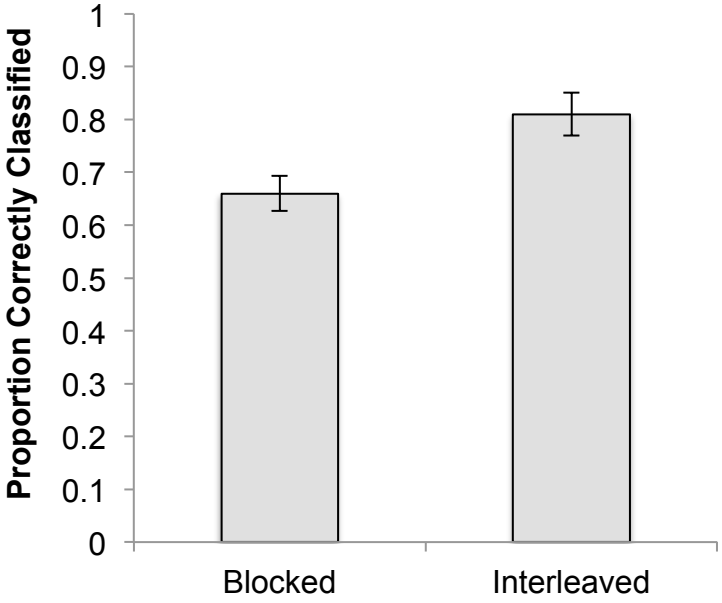

Figure 1. The proportion of new problems correctly classified on the final test as a function of study schedule, in Experiment 1. Error bars represent standard error of the mean.

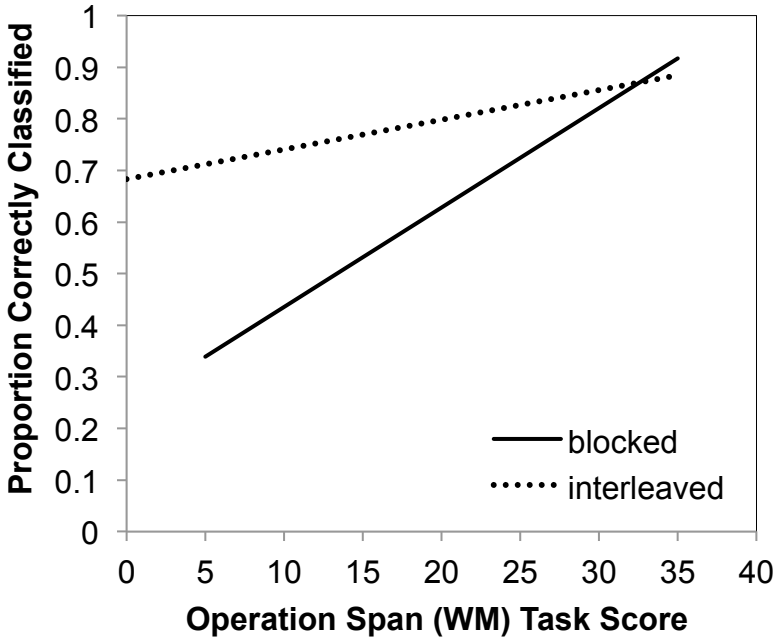

Figure 2. Linear regression slopes for classification performance and working memory scores as a function of study schedule, in Experiment 1.

Experiment 2

The goal of Experiment 2 was to examine the interaction between study schedules (blocked versus interleaved) and temporal spacing (spacing vs. no-spacing) on the inductive learning of statistical concepts. If the advantage of interleaving is due to discriminative contrast rather than spaced retrieval, then disrupting this discrimination process should decrease classification performance. Thus, when study problems are temporally spaced apart with unrelated fillers, this interleaving benefit may be eliminated because the filler disruption may offset the discrimination processing necessary to produce the benefit.

Results supporting this prediction are mixed. To test for discriminative-contrast hypothesis, studies have compared two interleaved conditions, one with and the other without temporal spacing. Whereas Birnbaum et al. (2013; butterfly species) found that adding spacing impaired learning presumably because the fillers interrupted the contrast processes, Zulkiply and Burt (2012) did not find such an effect as both conditions produced similar learning gains.

We also proposed a blocking benefit when study problems are temporally spaced apart, replicating the spacing effect. Again, results supporting this claim are mixed. Of the three studies that have compared blocked schedules with and without spacing, one reported a blocking benefit, consistent with the study-phase retrieval hypothesis (Birnbaum et al., 2013), but the other two did not find a blocking benefit (Kang & Pashler, 2012; Zulkiply & Burt, 2012).

We expected participants' WMC to predict their classification performance in the temporally-spaced interleaved study sequence. Lower WMC individuals encode contextual cues at a global rather than a specific level, which makes for a noisier search of target information at the time of retrieval (Kane & Engle, 2000; Rosen & Engle, 1998). Introducing temporal disruptions in an interleaved schedule, which may already produce some level of contextual interference, would further contaminate the cues (to include irrelevant features) that would be used to delimit the search set.

We did not expect participants' WMC to predict their classification performance

in the temporally-spaced blocked study sequence. In fact, we propose that the temporal spacing in the blocked schedule would particularly help lower WMC participants as the interval between problems would promote some forgetting of features, but the subsequent problem presentations would serve as reminders to retrieve those features again. This would allow the participants to encode more refined and relevant search cues as they accumulate several retrieval routes to access, and to strengthen memory traces of, a concept's critical features (Brewer & Unsworth, 2012; Tse & Pu, 2012).

Method

Participants. One hundred and thirty seven undergraduate students (94 females; *M* age = 18.31 years, *SD* = .99) from McMaster University participated in the experiment in exchange for 1 course credit. There were 35 participants each in the blocked-sequential condition, and 34 participants each in the interleaved-sequential condition and the blocked-spaced condition, and the interleaved-spaced condition. There were no significant differences in age or statistical background between the two conditions. In addition to the 137 participants, two participants completed the experiment but were excluded from the analysis for indicating previous knowledge of the statistical concepts being tested in the experiment.

Materials, Design, and Procedure. Participants were randomly assigned to one of four study conditions: blocked-sequential, interleaved-sequential, blocked-spaced, and interleaved-spaced. The blocked-sequential condition and interleaved-sequential condition were identical to the blocked and interleaved conditions, respectively, in Experiment 1; in the former, study problems were blocked by concept and in the latter, study problems cycled through the three concepts with no two problems of the same concept appearing consecutively. The blocked-spaced condition and interleaved-spaced conditions were identical to their sequential counterpart conditions with the exception that an unrelated 30-sec cartoon comic was inserted in between successive problem presentations. For all conditions, each study problem was accompanied with the concept's name directly above it.

Results and Discussion

Classification performance. We conducted a 2 (schedule: blocked vs. interleaved) x 2 (spacing: sequential vs. spaced) between-subjects ANOVA to examine the effects of schedule and spacing on classification performance. As illustrated in Figure 3, performance was significantly better in the interleaved conditions (M = .76. SD = .19) than in the blocked conditions (M = .69, SD = .20), as indicated by a main effect of schedule, $F(1,133) = 4.64$, MSE = .03, p = .033, $\eta\rho^2$ = .04. There was no main effect of spacing ($F < 1$). Of particular interest, however, was the significant interaction between schedule and spacing, $F(1, 133) = 8.80$, p = .004, $\eta\rho^2$ = .06.

Post-hoc comparisons revealed that when study problems were presented sequentially (without spacing), interleaved study (M = .81, SD = .18) produced better performance than blocked study (M = .65, SD = .19), $F(1, 133) = 13.20$, $p < .001$, $\eta\rho^2$ = .09. However, when spacing was inserted, there was no difference in performance between the blocked (M = .75, SD = .19) and interleaved schedules (M = .72, SD = .17; $F < 1$), consistent with the hypothesis that inserting fillers between exemplars prevents discrimination processing. In fact, with the spacing between study problems, participants in the blocked schedule numerically outperformed those in the interleaved schedule.

In line with our prediction, we also observed a marginally significant decrease in classification performance when spacing was added to the interleaved schedule, $F(1, 133) = 3.52$, p = .063, $\eta\rho^2$ = .03. Birnbaum et al. (2013) reported a similar, but significant decrement between the interleaved conditions when spacing was added. Together, this may be suggestive evidence for the role of between-concept comparisons in explaining the interleaving effect, as comparisons would be more difficult or disruptive in interleaved schedules with temporal spacing.

As predicted, in the blocked schedules, inserting spacing between study problems produced better performance than presenting problems consecutively (no-spacing), $F(1, 133) = 5.39$, p = .022, $\eta\rho^2$ = .04. This result is in line with the spacing effect literature (i.e., study-phase retrieval hypothesis; Cepeda et al., 2006, 2008), and consistent with the blocking benefit reported by Birnbaum et al. (2013). In the temporally-spaced blocked

schedule, the temporal spacing between problems of the same concept likely promoted forgetting, leading to more effortful retrieval, which strengthened the learning of that concept.

Individual differences. We regressed participants' WM scores on their classification performance for each schedule condition, as illustrated in Figure 4, to examine if WMC predicted performance in each of the four sequence conditions. As observed in Experiment 1, participants' WMC predicted their classification performance in the blocked-sequential condition, $t(1,33) = 4.21$, $p < .001$, $R^2 = .35$, but not in the interleaved-sequential condition, $t(1,32) = 1.89$, $p = .067$, $R^2 = .10$, which suggests that interleaving mitigates differences in cognitive ability.

Participants' WMC did not predict their classification performance in the blocked-spaced condition, $t(1, 32) = 1.67$, $p = .105$, $R^2 = .08$, suggesting that presenting study problems of a concept across time (i.e., the spacing effect; Cepeda, 2006, 2008, 2009) can decrease performance differences between higher and lower WMC individuals. Participants' WMC predicted their classification performance in the interleaved-spaced condition, $t(1,32) = 3.01$, $p = .005$, $R^2 = .22$, which was arguably the most disruptive and cognitively demanding condition.

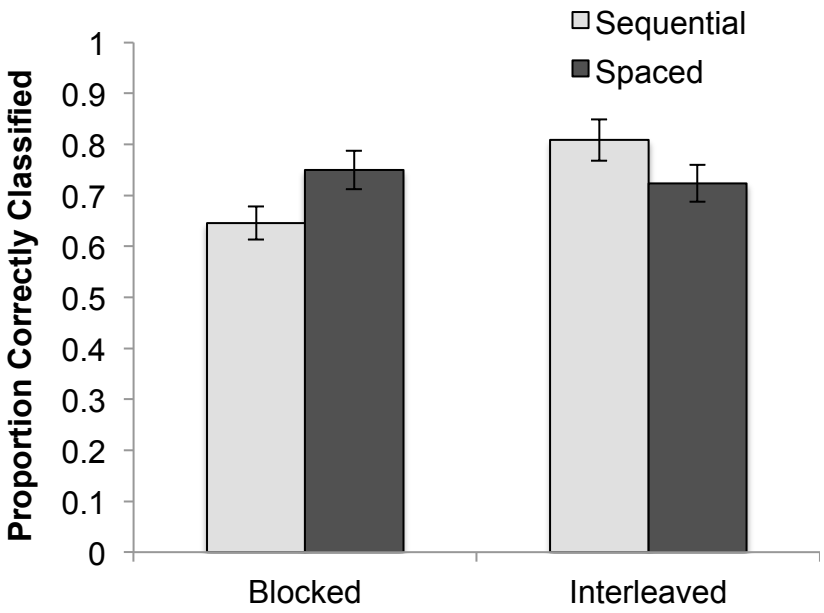

Figure 3. The proportion of new problems correctly classified on the final test as a function of study schedule, in Experiment 2. Error bars represent standard error of the mean.

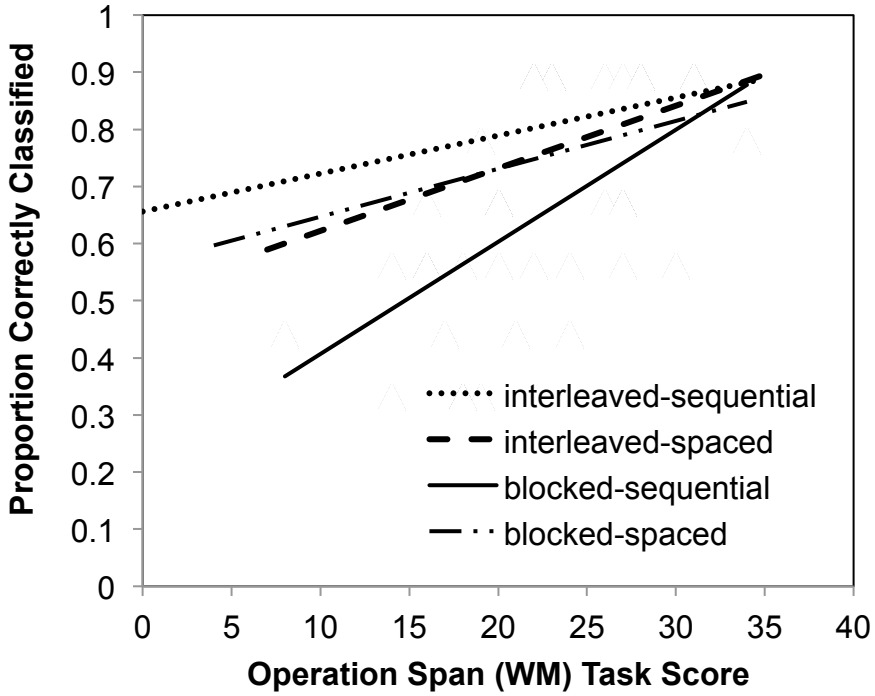

Figure 4. Linear regression slopes for classification performance and working memory scores as a function of study schedule and temporal spacing, in Experiment 2.

Experiment 3

The goal of Experiment 3 was to examine the interaction between study schedules (blocked versus interleaved schedules) and temporal juxtaposition (simultaneous vs. sequential sequences) on the inductive learning of statistical concepts. Study problems in Experiments 1 and 2 were presented one-at-a-time (i.e., sequentially). Viewing multiple problems of concepts at the same time may reduce the memory constraints, and provide a more explicit learning context to contrast the critical differences between concepts in the interleaved schedule, and to connect the critical commonalities of a given concept in the blocked schedule (Kang & Pashler, 2012; Mundy, Honey & Dwyer, 2007). Thus, we predict classification performance for concepts learned in simultaneous sequences to be better than for concepts learned in sequential sequences. This should hold true for blocked schedules (consistent with similarity-abstraction hypothesis) and for interleaved schedules (consistent with discriminative-contrast hypothesis).

The limited studies that examined the effect of simultaneous versus sequential study sequences on category learning under different study schedules have shown mixed results. When comparing interleaved schedules, Kang and Pashler (2012; artists' painting styles) found no learning benefit or impairment for simultaneous over sequential sequences, whereas Wahlheim et al. (2011; bird families) did find a learning benefit for simultaneous sequence. Perhaps with the relatively difficult stimuli of bird families (that produced lower test performance overall), the discriminative processes facilitated by interleaving can only be capitalized upon when the exemplars are presented simultaneously, which presumably decreases the memory load, and learners' attention is drawn more explicitly to the differences between the categories.

Neither of the studies discussed above showed benefits for simultaneous sequences in blocked schedules. On the other hand, studies on analogical transfer have shown learning to benefit when multiple items of the same category are compared simultaneously rather than studied separately (e.g., Catrambone & Holyoak, 1989; Gentner, Loewenstein, Thompson, & Forbus, 2009; Gick & Holyoak, 1983; Star & Rittle-Johnson, 2009). For instance, participants demonstrated greater learning of negotiation

principles when allowed to compare two case studies for a given negotiation strategy side-by-side rather than when these case studies were studied separately (Gentner et al., 2003). Simultaneously presenting problems fosters a more direct comparison, which in turn helps learners overcome contextual limitations and allows them to recognize the common deep features (e.g., Catrambone & Holyoak, 1989; Markman & Gentner, 2005). Moreover, problems that differ in their surface features (i.e., cover stories, storylines, events, names, objects) but share similar structural features (i.e., principles, equations, procedures), as is the case with our statistics stimuli, can further enable this comparison process as they more quickly realize what features are and what features are not relevant for categorization (e.g., Quilici & Mayer, 1996, 2002).

Finally, we expected that participants' WMC would not predict their classification performance in either of the simultaneous study sequences. This sequence would reduce memory demands, particularly helpful for lower WMC participants who may be more susceptible to irrelevant features (e.g., problems' cover stories) (Unsworth & Engle, 2007). Studying the problems three-at-a-time would allow them to allocate their attention to search for the relevant concept features, whether it is features shared by a concept or features different between concepts.

Method

Participants. One hundred thirty five undergraduate students (97 females; *M* age = 18.94 years, *SD* = 2.40) from McMaster University participated in the experiment in exchange for 1 course credit. There were 33 participants each in the blocked-sequential and interleaved-sequential conditions, 34 in the blocked-simultaneous condition, and 35 in the interleaved-simultaneous condition. There were no significant differences in age or statistical background between the two conditions.

Materials, design and procedure. The procedure for Experiment 3 was nearly identical to that of Experiment 1. The only difference was the addition of two more between-subject conditions (a total of four study conditions). Participants were randomly assigned to one of four study conditions: blocked-sequential, interleaved-sequential, blocked-simultaneous, and interleaved- simultaneous. The blocked-sequential condition

and interleaved-sequential condition were identical to the blocked and interleaved conditions in Experiment 1.

The blocked-simultaneous condition and interleaved-simultaneous conditions were identical to their sequential counterparts with the exception that the study problems were presented three-at-a-time (i.e., on the same page) for 60-sec, with a 3-sec blank screen after each set of three problems. In other words, total study time was held constant across the four conditions. In the blocked-simultaneous condition, all six exemplar problems from a given concept were presented before moving on to the next concept. These six exemplar problems were presented as two sets of three, simultaneously presented problems. In the interleaved-simultaneous condition, problems were also presented in sets of three, but each problem in a given set came from a different concept. For all conditions, each study problem was accompanied with the concept name directly above it.

Results and Discussion

Classification performance. We conducted a 2 (schedule: blocked vs. interleaved) x 2 (juxtaposition: sequential vs. simultaneous) ANOVA to examine the effects of schedule and juxtaposition on classification performance. Figure 5 illustrates performance on the classification test as a function of schedule and juxtaposition. Overall, the interleaved conditions (M = .81, SD = .18) yielded better performance than blocked conditions (M = .73, SD = .19), as indicated by a main effect of schedule, $F(1, 131) = 5.64$, MSE = .03, p = .019, $\eta\rho^2$ = .04. There was also a marginally significant main effect of juxtaposition, $F(1, 131) = 3.79$, MSE = .03, p = .054, η_ρ^2 = .03, with simultaneous sequences (M = .80, SD = .16) leading to better performance than sequential sequences (M = .74, SD = .20). In other words, presenting problems three-at-a-time, relative to presenting problems one-at-a-time, may have increased the extent to which participants noticed differences across problems of different concepts and noticed similarities across problems of the same concept.

The schedule x juxtaposition interaction was not significant (F<1). However, pairwise comparisons revealed that when study problems were presented sequentially,

interleaved schedule produced better performance than blocked schedule, $F(1, 131) = 5.57$, $p = .020$, $\eta_p^2 = .04$, replicating the interleaving effect. When study problems were presented simultaneously, interleaved and blocked schedules produced similar learning gains in classification ($F<1$, $p = .32$). This suggests that both schedules can be equally beneficial if the comparisons are made more explicit and if the memory demand is lower, as was the case with simultaneous sequences.

When schedules were interleaved, performance was better numerically, but not significantly so, for problems presented simultaneously rather than sequentially ($F<1$, $p = .54$), perhaps because performance was already at ceiling. This finding, nonetheless, may provide suggestive evidence for simultaneous sequences encouraging between-concept comparisons.

When schedules were blocked, simultaneously presenting study problems yielded better performance than sequentially presenting study problems, $F(1, 131) = 4.31$, $p = .040$, $\eta_p^2 = .03$. This finding is inconsistent with the category induction findings that show a lack of superior performance for blocked schedules with simultaneous sequences (Kang & Pashler, 2012; Wahlheim et al., 2011). These studies, however, use perceptual stimuli, which may drive induction differently. Research on learning with text-based stimuli suggest that learning is better when problems are designed to allow for useful inferences with regard to their structural and surface features (e.g., Gick & Holyoak, 1987; Quilici & Mayer, 1996). In the current experiment, we presented participants with three problems at once, each with different surface features, which may have promoted inter-problem processing (i.e., focusing on common features) over intra-problem processing (i.e., focusing on specific wording or details; e.g., Gentner et al., 2003; Gick & Holyoak, 1983; Quilici & Mayer, 1996, 2002).

Individual differences. Figures 6 shows the linear regression lines for each study sequence as a function of participants' WMC on their classification test performance. Consistent with findings from Experiment 1 and 2, participants' WMC predicted their classification performance in the blocked-sequential condition, $t(1,31) = 3.95$, $p < .001$, $R^2 = .34$, but not in the interleaved-sequential condition, $t < 1$, $p = .44$, $R^2 = .02$.

Participants' WMC did not predict their classification performance in the interleaved-simultaneous condition, $t(1,33) < 1$, $R^2 = .0$, suggesting that juxtaposing problems of different concepts, three-at-a-time or one-at-a-time, decreases performance differences across lower and higher WMC individuals. Surprisingly, participants' WMC predicted their classification performance in the blocked-simultaneous condition, $t(1,32) = 2.69$, $p = .011$, $R^2 = .19$. Note that the correlation .19 is lower than .34 observed in the blocked-sequential condition, suggesting that although performance differences between lower and higher WMC participants did decrease in the blocked-simultaneous condition, this learning benefit may not be comparable to the learning benefit conferred with an interleaved schedule.

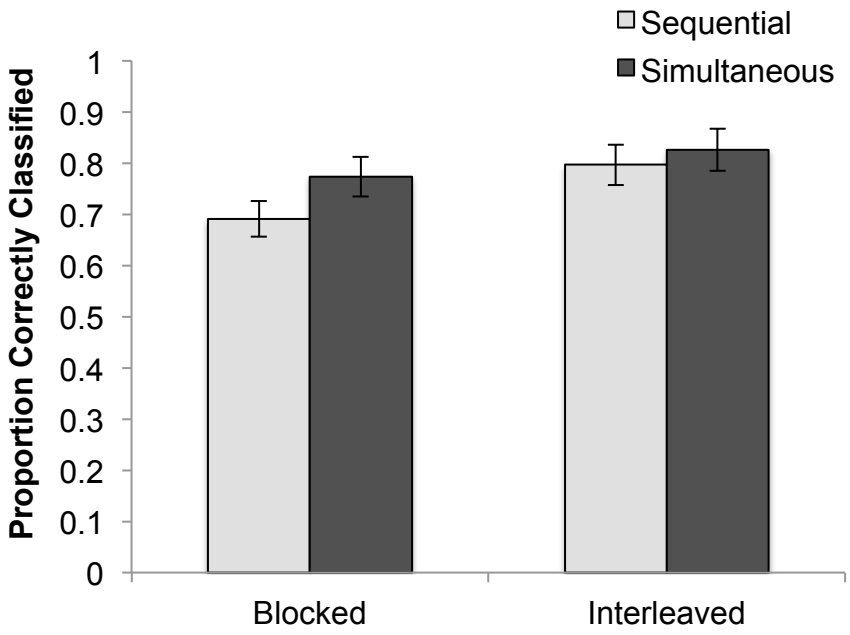

Figure 5. The proportion of new problems correctly classified on the final test as a function of study schedule, in Experiment 3. Error bars represent standard error of the mean.

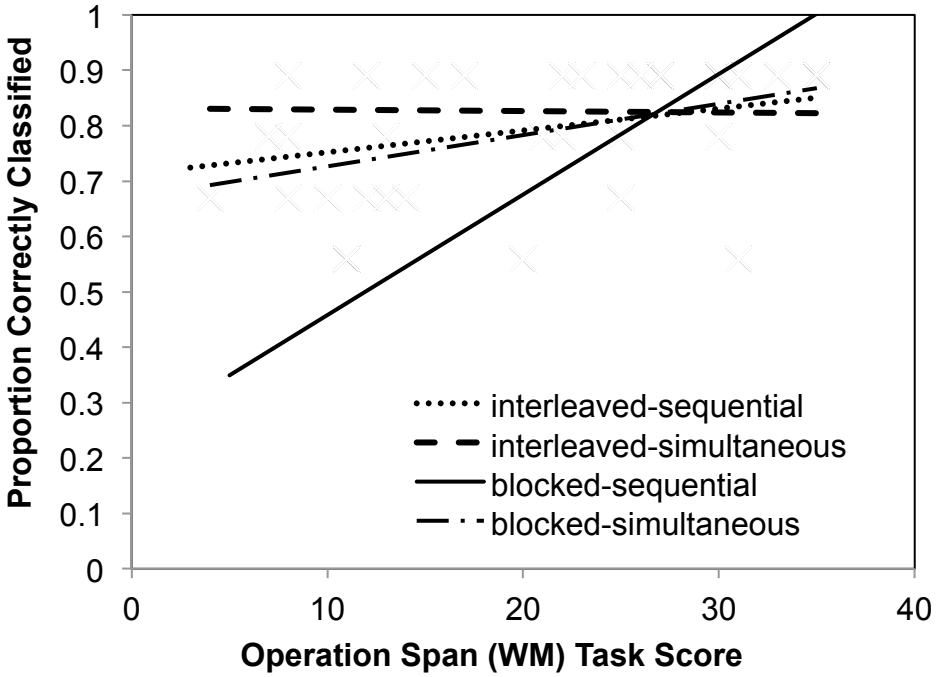

Figure 6. Linear regression slopes for classification performance and working memory scores as a function of study schedule and juxtaposition, in Experiment 3.

General Discussion

At a basic level, a statistics course consists of several highly abstract concepts that the students must learn. While many may learn the procedural skills necessary to correctly calculate test concepts, students often struggle with the conceptual knowledge of *when* to correctly apply the procedure. This issue is exacerbated by the fact that most textbooks emphasize practice in using procedures but not in determining which procedures to use, and when to use them (Mayer, Sims, & Tajika, 1995). To acquire conceptual knowledge, one needs to build a cognitive structure that represents an understanding of statistics in terms of links and relations between the important concepts. Our studies demonstrate that this conceptual learning can be accomplished by juxtaposing study problems of different concepts, or by studying problems of the same concept distributed across time. Furthermore, such instructional strategies are especially beneficial for learners with lower cognitive abilities.

Classification performance. Findings from Experiment 1 demonstrate that the

interleaving benefit generalizes to the inductive learning of statistical concepts. Participants were better able to identify the concepts of previously unseen test problems when study problems were interleaved with those of other concepts rather than blocked by concept.

Results of Experiment 2 demonstrate an interaction between study schedules and temporal spacing: increasing spacing by inserting unrelated fillers between successive study problems decreased classification performance in the interleaved schedule, consistent with the discriminative-contrast hypothesis, and increased classification performance in the blocked schedule, consistent with the study-phase retrieval hypothesis. Participants were better at classifying test problems when study problems in the interleaved schedule were not temporally spaced apart—the unrelated fillers presumably disrupted the contrast processes critical to between-concept comparisons. Participants were also better at classifying test problems when study problems in the blocked schedule were temporally spaced apart—learning was better because the temporal delay between problems allowed time to forget the features, which made retrieval of those features from memory somewhat difficult, but successful.

Results from Experiment 3 suggest that simultaneous study sequences improved overall classification performance, particularly in the blocked schedule, consistent with the commonality-abstraction hypothesis. Participants were better at classifying test problems when study problems in the blocked schedule were presented three-at-a-time rather than one-at-a-time. A simultaneous study sequence may reduce memory load to provide a more explicit learning context to elicit the critical commonalities within a concept, and to extract the relevant features shared across problems, and disregard the irrelevant surface features. Participants did not differ in their ability to classify test problems when study problems in the interleaved schedule were presented three-at-a-time rather than one-at-a-time (although the former scored higher than the latter numerically). We do not know if this lack of a significant difference is due to the classification performance being at ceiling or due to a real effect (or lack thereof).

The finding that both increasing temporal spacing (Experiment 2) and eliminating

temporal spacing (Experiment 3) produced greater learning gains, at first glance, may appear to contradict one another, but suggests that temporal factors may recruit different cognitive processes. Increasing spacing enhances learning by encouraging forgetting and subsequently retrieving concept features from memory, whereas decreasing (or eliminating) spacing enhances learning by encouraging greater comparison of and attention to common, relevant features. An interesting follow-up question would be to examine whether the learning gains from spaced effortful retrieval are longer lasting than the learning gains from explicit problem comparisons.

Findings from the perceptual category induction research on study schedules contradict some of the findings in the current study. Specifically, when comparing blocked schedules, we found a spacing advantage, whereas Kang & Pashler (2012) and Zulkiply and Burt (2012) did not. When comparing interleaved schedules, we and Birnbaum et al. (2013), found a spacing disadvantage, whereas Zulkiply and Burt (2012) did not. And finally, when comparing blocked schedules, we found an advantage for simultaneous study sequence, whereas Kang and Pashler (2012) and Wahlheim et al. (2011) did not. With the exception of Birnbaum et al. (2013) who examined the inductive learning of butterfly species, the other studies focused on the inductive learning of artists' painting styles (Kang & Pashler, 2012; Zulkiply & Burt, 2012) and of bird species (Wahlheim et al., 2011). Perhaps the different pattern of results may be because, unlike learning artists' painting styles, characteristics that define different butterflies are verbalizable (e.g., specific patterns, shapes and colours of the wings), similar to the rule-based statistical concepts. There is some evidence to suggest that blocked schedules encourage explicit hypothesis testing, which is more favourable to rule-based or feature-based categories and concepts. However, this assertion merits further empirical exploration.

Individual differences in WMC. We also examined the relations between participants' cognitive abilities and their classification performance as a function of the different study sequences. Across the three experiments, WMC predicted classification performance when the study sequences were blocked-sequential, blocked-simultaneous

and interleaved-spaced. On the other hand, WMC did not predict classification performance when the study sequences were interleaved-sequential, interleaved-simultaneous and blocked-spaced.

Interestingly, the observed learning detriments and gains seem to be driven particularly by participants with lower WMC. This result is not surprising as individuals with higher WMC are better at controlling their attention to process task relevant information, and controlling their search of LTM to retrieve relevant information (Unsworth & Engle, 2007). Moreover, these individuals may already use efficient strategies to process information unlike their lower WMC counterparts (e.g., Brewer & Unsworth, 2012), and therefore, they may be less susceptible to the different study sequence manipulations.

There are several theoretical possibilities that may explain why lower WMC individuals may be more susceptible to specific sequence manipulations. The extent to which individuals successfully retrieve encoded information from LTM depends on their ability to engage in controlled and strategic cue-dependent search of LTM (Unsworth & Engle, 2007). Low-WMC individuals encode contextual cues at a global level rather than a specific level, and therefore, their search set is often noisier, which increases intrusions and errors (Kane & Engle, 2000; Rosen & Engle, 1998).

As proposed by the study-phase retrieval hypothesis, subsequent problem presentations of the same concept in interleaved schedules, and a blocked schedule with temporal spacing served as reminders to retrieve previously studied features again. The forgetting and retrieval allowed lower WMC participants to accumulate several retrieval routes to access, and to strengthen memory traces of, the concept's critical features, which ultimately constrained their search set to include more diagnostic features that repeatedly occur across subsequent problems of a concept (Brewer & Unsworth, 2012; Tse & Pu, 2012). Conversely, the interleaved schedule with temporal spacing, a sequence that already produces high levels of contextual interference, may have promoted unsuccessful retrieval attempts of previously studied features, and included irrelevant or indistinguishable cues used to delimit the search set, which ultimately impaired concept

learning.

As for the simultaneous study sequences, the enhanced discriminative contrast and commonality abstraction associated with interleaving and blocking may have increased the salience of relevant features between and within concepts, which possibly alleviated limitations to concept induction associated with having low WMC; individuals with lower WMC less likely to resist interference caused by irrelevant information, such as cover stories of word-problems that are not relevant to the concept learning.

Further studies are warranted to replicate and more closely investigate the relation between study sequences and individual differences in WMC, and how these variables may interact with each other. Findings from the current study demonstrate that interleaved schedules and blocked-temporally-spaced schedule can decrease learning differences across individuals with varying cognitive abilities, and increase overall learning for all individuals, suggesting that these sequences are optimal for the inductive learning of statistical concepts.

Concluding Remarks

We examined the different factors and processes that determine *when* and *how* one schedule may be more or less effective than the other. Interleaved schedules are effective when problems of different concepts are presented (together or individually) with no disruptions in-between the problems. This seems to be because interleaving facilitates between-concept comparison that is susceptible to disruptions. Blocked schedules are effective when problems of the same concept are juxtaposed three-at-a-time, as it provides a contextual learning environment that allows learners to make explicit comparisons and extract the relevant features shared by all problems from the irrelevant features. Blocked schedules are also effective when there is some form of temporal spacing introduced in-between problems, as the temporal lag allows for some forgetting and subsequent retrieval of problem features, memory traces for which are then strengthened.

Properly constructed study sequences can promote concept learning in the domain of statistics. Our use of ecologically valid materials offers some confidence in the

generalizability of these findings to other cognitive concepts. There are several straightforward practical ways in which our findings can be implemented in formal educational settings. Course assignments and end-of-chapter practice problems in textbooks may include problems not just from the current topic, but also from previous units or chapters. Instructors can also lead explicit discussions on the commonalities and differences across concepts. This direct instruction may offer benefits not only in terms of increasing students' understanding, but also in terms of encouraging students to search out comparison opportunities on their own.

References

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. doi: 10.1016/j.jecp.2009.11.003

Baddeley, A.D., & Hitch, G.J. (1974). Working memory. In G. H. Bower (Ed.), *The psychology of learning and motivation, Vol. 8* (pp. 47-89). New York: Academic Press.

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. doi: 10.3758/s13421-012-0272-7

Bjork, R.A. (1975). Retrieval as a memory modifier. In R. Solso (Ed.), *Information processing and cognition: The Loyola Symposium* (pp. 123-144). Hillsdale, NJ: Lawrence Erlbaum Associates.

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

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

Carvalho, P.F. & Goldstone, R.L. (2015). What you learn is more than what you see: What can sequence effects tell us about inductive category learning? *Frontiers in Psychology, 6*(505). doi: 10.3389/fpsyg.2015.00505

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

Catrambone, R., & Holyoak, K.J. (1989). Overcoming contextual limitations on problem-solving transfer. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, *15*(6), 1147–1156. doi: 10.1037/0278-7393.15.6.1147

Cepeda, N.J., Coburn, N., Rohrer, D., Wixted, J.T., Mozer, M.C., & Pashler, H. (2009).

Optimizing distributed practice: theoretical analysis and practical implications. *Experimental Psychology, 56*(4), 236–246. doi: 0.1027/1618-3169.56.4.236

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

Cepeda, N.J., Vul, E., Rohrer, D., Wixted, J.T., & Pashler, H. (2008). Spacing effects in learning: a temporal ridgeline of optimal retention. *Psychological Science, 19*(11)*,* 1095–1102. doi: 10.1111/j.1467-9280.2008.02209.x

Dempster, F.N. (1988). The spacing effect: A case study in the failure to apply the results of psychological research. *American Psychologist*, *43*(8), 627-634. doi: 10.1037/0003-066X.43.8.627

Engle, R.W., & Kane, M.J. (2004). Executive attention, working memory capacity, and a two-factor theory of cognitive control. In B. Ross (Ed.), *The psychology of learning and motivation, Vol. 44* (pp. 145-199). New York: Elsevier.

Gentner, D. (1983). Structure-mapping: A theoretical framework for analogy. *Cognitive Science*, *7*(2)*,* 155–170. doi: 10.1207/s15516709cog0702_3

Gentner, D., Loewenstein, J., & Thompson L. (2003). Learning and transfer: A general role for analogical encoding. *Journal of Educational Psychology, 95*(2), 393– 408. doi: 10.1037/0022-0663.95.2.393

Gentner, D., Loewenstein, J., Thompson, L., & Forbus, K.D. (2009). Reviving inert knowledge: Analogical abstraction supports relational retrieval of past events. *Cognitive Science*, *33*(8), 1343-1382. doi: 10.1111/j.1551-6709.2009.01070.x

Gick, M.L., & Holyoak, K.J. (1983). Schema induction and analogical transfer. *Cognitive Psychology, 15*(1)*,* 1-38. doi: 10.1016/0010-0285(83)90002-6

Gick, M.L., & Holyoak, K.J. (1987). The cognitive basis of knowledge transfer. In S. M. Cormier & J. D. Hagman (Eds.), *Transfer of learning: Contemporary research and applications* (pp. 9-46). Orlando, FL: Academic Press.

Goldstone, R.L. (1996). Isolated and interrelated concepts. *Memory & Cognition. 24*(5)*,* 608–628. doi: 10.3758/BF03201087

Gravetter, F.J., & Wallnau, L.B. (2008) *Statistics for the behavioral sciences* (8th ed.). Belmont, CA: Wadsworth Cengage Learning.

Kane, M.J., & Engle, R.W. (2000). Working-memory capacity, proactive interference, and divided attention: Limits on long-term memory retrieval. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, *26*(2), 336-358. doi: 10.1037/0278-7393.26.2.336

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

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

Kornell, N., Castel, A.D., Eich, T.S., & Bjork, R.A. (2010). Spacing as the friend of both memory and induction in young and older adults. *Psychology & Aging*, *25*(2), 498-503. doi: 10.1037/a0017807

Kurtz, K.H., and Hovland, C.I. (1956). Concept learning with differing sequences of instances. *Journal of Experimental Psychology, 51*(4), 239–243. doi: 10.1037/h0040295

Kurtz, K.J., Miao, C.H., & Gentner, D. (2001). Learning by analogical bootstrapping. *The Journal of the Learning Sciences*, *10*(4), 417-446. doi: 10.1207/S15327809JLS1004new_2

Le Blanc, K., & Simon, D. (2008, November). Mixed practice enhances retention and JOL accuracy for mathematical skills. In *49th Annual Meeting of the Psychonomic Society, Chicago, IL*.

Markman, A.B., & Gentner, D. (2005). Nonintentional similarity processing. In R. Hassin, J.A. Bargh and J.S. Uleman (Eds.), *The new unconscious* (pp. 107-137). New York: Oxford University Press.

Mayer, R.E., Sims, V., & Tajika, H. (1995). A comparison of how textbooks teach mathematical problem solving in Japan and the United States. *American*

Educational Research Journal, 32(2)*,* 443–460. doi: 10.3102/00028312032002443

Mayfield, K.H., & Chase, P.N. (2002). The effects of cumulative practice on mathematics problem solving. *Journal of Applied Behavior Analysis*, *35*(2), 105-123. doi: 10.1901/jaba.2002.35-105

Mundy, M.E., Honey, R.C., and Dwyer, D.M. (2007). Simultaneous presentation of similar stimuli produces perceptual learning in human picture processing. *Journal of Experimental Psychology: Animal Behavior Processes, 33*(2), 124–138. doi: 10.1037/0097- 7403.33.2.124

Quilici, J.L., & Mayer, R.E. (1996). Role of examples in how students learn to categorize statistics word problems. *Journal of Educational Psychology*, *88*(1), 144-161. doi: 10.1037/0022-0663.88.1.144

Quilici, J.L., & Mayer, R.E. (2002). Teaching students to recognize structural similarities between statistics word problems. *Applied Cognitive Psychology*, *16*(3), 325-342. doi: 10.1002/acp.796

Rohrer, D. (2009). The effects of spacing and mixing practice problems. *Journal for Research in Mathematics Education, 40*(1), 4-17. Retrieved from http://www.jstor.org/stable/40539318

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

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

Rosen, V.M., & Engle, R.W. (1998). Working memory capacity and suppression. *Journal of Memory & Language*, *39*(3), 418-436. doi: 10.1006/jmla.1998.2590

Star, J.R., & Rittle-Johnson, B. (2009). It pays to compare: An experimental study on computational estimation. *Journal of Experimental Child Psychology*, *102*(4), 408-426. doi: 10.1016/j.jecp.2008.11.004

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

Thios, S.J., & D'Agostino, P.R. (1976). Effects of repetition as a function of study-phase retrieval. *Journal of Verbal Learning & Verbal Behavior*, *15*(5), 529-536. doi: 10.1016/0022-5371(76)90047-5

Tse, C.S., & Pu, X. (2012). The effectiveness of test-enhanced learning depends on trait test anxiety and working memory capacity. *Journal of Experimental Psychology: Applied*, *18*(3), 253-264. doi: 10.1037/a0029190

Unsworth, N., & Engle, R.W. (2007). The nature of individual differences in working memory capacity: active maintenance in primary memory and controlled search from secondary memory. *Psychological Review*, *114*(1), 104-132. doi: 10.1037/0033-295X.114.1.104

Unsworth, N., & Spillers, G.J. (2010). Variation in working memory capacity and episodic recall: The contributions of strategic encoding and contextual retrieval. *Psychonomic Bulletin & Review*, *17*(2), 200-205. doi: 10.3758/PBR.17.2.200

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. doi: 10.3758/BF03192720

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

Whitman, J.R., & Garner, W.R. (1963). Concept learning as a function of form of internal structure. *Journal of Verbal Learning & Verbal Behavior*, *2*(2), 195-202. doi: 10.1016/S0022-5371(63)80085-7

Zulkiply, N., & Burt, J.S. (2012). The exemplar interleaving effect in inductive learning: Moderation by the difficulty of category discriminations. *Memory & Cognition*, *41*(1), 16-27. doi: 10.3758/s13421-012-0238-9

Chapter 4: Learning Categories from Exemplars: Does the Optimal Schedule of Presentation Vary as a Function of Within-Category Versus Between-Category Similarity of Exemplars?

Study Motivation and Overview

In this chapter, I examined two factors: category similarity structure and the nature of the to-be-learned categories that can potentially moderate the interleaving effect. While in the previous chapter, I made the process of making discriminations and comparisons more or less difficult via temporal spacing, in this chapter, I varied the importance of making between-category discriminations by manipulating category similarity structure and examined the interaction between this structure manipulation and study schedules. Carvalho and Goldstone (2014) manipulated within- and between-category similarity together, and Zulkiply and Burt (2012) focused only on between-category discriminations. Even previous studies that used realistic or naturalistic materials, such as artists' painting styles, consisted only of landscape paintings which were constructed with both high within- and between-category similarity together (Carvalho & Goldstone, 2014, Experiment 1; Kang & Pashler, 2012; Kornell & Bjork, 2008). In order to directly test the hypotheses that a blocked study schedule fosters the noticing of within-category commonalities and an interleaved study schedule fosters the noticing of between-category discriminations, I examined category structures with high within- and low between-category similarity and the inverse, low within- and high-between category similarity, where similarity is defined by irrelevant characteristics of the stimuli (e.g., the surface "cover story" of the research designs or the subjects of the paintings). By making the commonalities or the differences trivially easy to spot (e.g., all of the research designs appropriately analyzed using a chi-square test concerned cover stories about schooling, or all of the paintings by the artist Grossman consisted of flowers), these hypotheses make very clear predictions for when blocking is beneficial and when interleaving is beneficial.

I also examined whether the optimal study schedules are similar or different depending the nature of the categories that are to be learned. I compared perceptual-

based categories (i.e., artists' painting styles) with conceptual-based categories (i.e., statistical concepts). Compared to categories of artistic styles that are perceptually rich, categories of non-parametric statistical tests rely less on perceptual features and more on structural features and rules to abstract commonalities within a category or to distinguish subtle features between categories. Thus, we may see a different pattern of results given that "information-integration" and "rule-based" categories may have different optimal schedules.

Learning Categories from Exemplars: Does the Optimal Schedule of Presentation Vary as a Function of Within-Category Versus Between-Category Similarity of Exemplars?

Veronica X. Yan[1], Faria Sana[2], Joseph A. Kim[2], Elizabeth Ligon Bjork, and Robert A. Bjork[1]

Author Note

[1]Department of Psychology, University of California, Los Angeles, 1285 Franz Hall, Los Angeles, CA 90095-1563, United States.

[2] Department of Psychology, Neuroscience, & Behaviour, McMaster University, 1280 Main Street West, Hamilton, ON L8S 4K1, Canada.

Acknowledgments

This research was supported by Grant No. 29192G from the James S. McDonnell Foundation and by 767-2012-2053 Scholarship from Social Sciences and Humanities Research Council. Many thanks to the members of CogFog for their feedback.

Abstract

Interleaving exemplars of to-be-learned categories, rather than blocking exemplars by category, frequently enhances category induction. Recent findings, however, suggest that category similarity structure modulates this interleaving benefit, such that interleaving is more or less effective than blocking depending on whether successful categorization depends more on discriminating between exemplars of different categories or encoding commonalities across exemplars within a category (e.g., Carvalho & Goldstone, 2014; Zulkiply & Burt, 2012). In four experiments, we examined both within-category and between-category similarity of studied exemplars and we did so for two types of educationally realistic categories, one that is relatively rule-based (non-parametric statistical concepts, Experiments 1 and 3A) and one that is relatively perception-based (different artists' painting styles, Experiments 2, 3B, and 4). When between-category similarity was high, learning of both category types benefited from interleaving; but when between-category similarity was low, this interleaving benefit was eliminated for the perception-based categories, and a blocking benefit was obtained for the rule-based categories. Overall, these findings are consistent with the argument that the benefit of interleaving is modulated by within-category versus between-category similarity of studied exemplars and by type of category as well (i.e., perception-based versus rule-based).

Keywords: Interleaving, category induction, comparisons, contrasts, learning

Introduction

Much of learning—both in childhood and throughout the lifetime—is comprised of the inductive learning from exemplars of concepts and categories, which then leads to the ability to categorize new exemplars when they are encountered in the same or different situations. For instance, the clinical psychologist who has examined multiple cases of bipolar depression, and thus can easily identify the key symptoms defining this disorder, should be able to recognize not only a new case of bipolar depression, but should also be able to differentiate that case from cases of other subcategories of depression. This type of category knowledge requires awareness of both the similarities between instances of the same category and the differences between instances from different categories. A clinical psychology student may choose to study several cases of bipolar depression consecutively to extract the core symptoms of the disorder (i.e., make within-category comparisons), or to mix together the study of several different cases of subcategories of depression (e.g., bipolar, major, psychotic, postpartum) in order to learn the subtle differences that exist between them (i.e., make between-category comparisons). Each study method can be useful depending on whether it is more important to learn similarities within several cases of bipolar depression or differences between the various types of depression.

Indeed, evidence suggests that juxtaposing study of exemplars from the same category (i.e., blocking) fosters within-category comparisons (e.g., Gentner & Namy, 1999; Gentner, Loewenstein, Thompson, & Forbus, 2009; Oakes & Ribar, 2005), whereas juxtaposing exemplars from different categories (i.e., interleaving) fosters between-category comparisons (e.g., the discriminative contrast hypothesis; Birnbaum, Kornell, Bjork, & Bjork, 2013; Kang & Pashler, 2012; Kornell & Bjork, 2008). Each of these two very different schedules—blocked and interleaved—foster unique category comparisons, and one schedule may be more effective than the other depending on the category similarity structures. The main purpose of the present paper, therefore, is to examine the possible interaction of two factors—category similarity structure (i.e., the within- and between-category similarities among the set of categories to be learned) and presentation

schedule (interleaved and blocked study)—in the learning of realistic and educationally relevant categories. Furthermore, we examined the effects of category similarity structure and schedule on two very different types of categories, relatively conceptual (statistics) and perceptual (artists' painting styles) categories.

Prior Relevant Research

Findings from early research in the domain of category learning have tended to favor blocked over interleaved schedules of learning. Gagné (1950), for example, found that blocking nonsense-form categories led to better performance and fewer errors during the last two trials of acquisition than did interleaving those categories. Similarly, Kurtz and Hovland (1956) presented participants with four categories of geometric patterns, which varied along four relevant dimensions (shape, color, size, and position; a given category was defined by two of these four dimensions) and one irrelevant dimension. Although they found no performance differences between blocked and interleaved study in the classification of exemplars during acquisition, blocking led to better verbalization of the category-defining rules. More recently, Carpenter and Mueller (2013) found that French pronunciation rules regarding word endings (e.g., *–ir, –on* and *–is)* were learned equally well whether studied in a blocked or interleaved manner, but that blocking led to better learning of the pronunciation of whole words (as opposed to just the correct pronunciation of the rule portion of the word), perhaps because blocking the word endings allowed learners to focus on learning about how to pronounce word stems.

In contrast to these findings, a large body of research has found interleaved study to leads to better categorization performance than blocked study, particularly for the learning of complex, perceptual-based categories. Kornell and Bjork (2008) first demonstrated this effect for the learning of different artists' painting styles. Their participants studied six paintings by each of 12 artists, either blocked by artist (i.e., the six paintings of a given artist were presented consecutively) or interleaved with the paintings of other artists (i.e., presentation of the six paintings of a given artist were separated by presentations of paintings from the other artists whose styles were also to be learned). During a final test in which participants were asked to classify never-before-seen

paintings by the studied artists (i.e., identify which of the studied artist was responsible for the presented painting), participants' classification performance was better for artists whose paintings had been studied in an interleaved manner versus a blocked manner. Nevertheless, after the final test, participants overwhelmingly reported thinking that blocking had been better than interleaving for their learning of the different artists' styles.

Since Kornell and Bjork's study, the interleaving benefit has been replicated for older adults (Kornell, Castel, Eich, & Bjork, 2010) and with paintings of different artists (Kang & Pashler, 2012), bird families (Birnbaum et al., 2013; Wahlheim, Dunlosky, & Jacoby, 2011), and butterfly species (Birnbaum et al., 2013). Furthermore, the observed interleaving benefit has not been limited to the learning of perceptual-based categories. Zulkiply, McLean, Burt, and Bath (2012), for example, have found that participants could better identify psychopathology disorders in new case studies following interleaved, rather than blocked, presentations of exemplars (whether presented vocally or visually) during acquisition. Other studies have demonstrated that students are better able to discriminate among different types of math problems and apply the appropriate formula on a delayed test when their initial practice interleaved study of different problem types rather than blocking practice by problem type (e.g., Taylor & Rohrer, 2010; Rohrer & Taylor, 2007).

Whereas the studies described above focused on the question of *which* type of presentation schedule—interleaving or blocking—was more effective for category learning, other studies have focused on identifying *when* each type of schedule might be more effective for category learning. One factor that appears to affect schedule efficacy is category similarity structure—that is, the similarity relations within and between the categories in a given set of to-be-learned categories. Findings from this type of research indicate that interleaving (which is assumed to promote between-category comparison) is more effective for low-discriminability categories—that is, where all categories are very similar to one another; whereas, blocking (which is assumed to promote within-category comparison) is more effective for high-discriminability categories—that is, where all the

exemplars from the same category are highly dissimilar and the few features they share are difficult to identify.

Such a pattern of results was, in fact, observed by Carvalho and Goldstone (2014) and by Zulkiply and Burt (2012) when they manipulated category similarity structure so as to make either the between-category or the within-category similarities more difficult to learn. Carvalho and Goldstone (2014) created blob-shaped categories (defined by a particular notch in one location of the blob) whose exemplars shared few similarities within and between categories (low similarity categories) and found that blocking (where categories alternated only 25% of the time) produced better subsequent classification performance than did interleaving (where categories alternated 75% of the time). This pattern of results was reversed for blob-shaped categories whose exemplars shared a high level of similarity with other exemplars in the same category as well as with exemplars in different categories (high similarity categories). Zulkiply and Burt (2012), who manipulated difficulty of category discriminations, found similar results: namely, that blocking was more effective for learning highly discriminable categories (with presumably low between-category similarity), whereas interleaving was critical for inducing learning of low-discriminability categories (with presumably high between-category similarity).

Overview of Current Studies and Predictions

The goal of the present research was to examine more fully the interactions between presentation schedule (interleaved and blocked study) and category similarity structure (i.e., the within- and between-category similarities of the set of to-be-learned categories) in influencing the effectiveness of category learning. Specifically, in four experiments, we investigated whether the type of relation between schedule and category similarity structure observed in the previously described studies of Carvalho and Goldstone (2014) and Zulkiply and Burt (2012) would (a) extend to the learning of more educationally realistic categories and (b) whether this relation would be similar across two very different types of categories: conceptual, feature-defined or rule-based categories (e.g., non-parametric statistical tests) and perceptual-based, less easily

verbalizable categories (e.g., artists' painting styles). The conceptual categories we used, like many educationally relevant categories, relied less on perceptual features and more on structural features and rules to abstract commonalities within a category or to distinguish subtle features between categories. Similar to the stimuli used by Carvalho and Goldstone (2014) and by Zulkiply and Burt (2012), however, there were certain features that defined when a particular non-parametric statistical test should be used. On the other hand, the perceptual-based categories we used, like many real-life categories, relied more on flexible combinations of brushstroke technique, color palette, and content matter—thus categories that would be difficult to define in terms of a rule or even a set of rules.

The theoretical framework guiding the present research was that interleaving emphasizes differences among exemplars between different categories, whereas blocking emphasizes similarities among exemplars within the same category (e.g., Carvalho & Goldstone, 2014; Zulkiply & Burt, 2012). In line with this framework, we manipulated the similarity relations within and between categories depending on whether we wanted our stimuli to shift the learner's attention to the processing of differences or similarities. More specifically, in the first of the present experiments, we created sets of to-be-learned categories that had different category similarity structures by manipulating the content of the exemplars within different sets of to-be-learned categories. Specifically, in one set of to-be-learned categories, we made the similarity of the exemplars within each category to be high in terms of their content and the similarity between the exemplars of the different categories to be low in terms of their content. We henceforth refer to the set of categories with this type of high-within and low-between similarity construction as *distinct* categories. In another set of to-be-learned categories, we made the similarity of the exemplars within each category to be low and the similarity between the exemplars of different categories to be high. We henceforth refer to the set of categories with this type of low-within and high-between similarity construction as *overlapped* categories. The difference between our sets of distinct and overlapped categories based on their content and their consequent category similarity structure is illustrated in Table 1.

If the learning processes invoked by these two types of category similarity structures are similar to those invoked by the artificial categories used by Carvalho and Goldstone (2014) and by Zulkiply and Burt (2012), we would expect to see a benefit of interleaving (assumed to foster between-category comparisons) in the learning of the overlapped categories, but a benefit of blocking (assumed to foster within-category comparisons) in the learning of the distinct categories. Findings consistent with this prediction would provide converging evidence for the influence of category similarity structure in determining schedule efficacy across the learning of different types of categories (e.g., artificial, naturalistic, and educationally relevant).

Is it reasonable, however, to make such a prediction—that is, to expect that this pattern of results would hold for the learning of conceptual types of categories, such as the different non-parametric statistical tests used in the present Experiment 1, and for perceptual types of categories, such as the different artists' painting styles used in the present Experiment 2? Compared to categories of artists' styles, categories of non-parametric statistical tests could be seen as more feature- or rule-based. Perhaps then, a more reasonable pattern of results to expect would be one based on the assumptions proposed by Ashby and colleagues (Ashby, Alfonso-Reese, Turken, & Waldron, 1998; Ashby & Maddox, 2005, 2010) who have drawn a distinction between "rule-based" category learning (i.e., the learning of categories that can be defined by easily verbalizable rules) and "information-integration" category learning (i.e., the learning of categories that cannot be defined by easily verbalizable rules). Two different systems are presumed to underlie rule-based and information-integration based category learning (an explicit and an implicit learning system, respectively; Ashby et al., 1998). Additionally, because blocked presentations of exemplars is thought to promote the search for rules and hypothesis testing, it is presumed to be more compatible for the learning of rule-based categories than for information- integration-based categories, the learning of which could even be impaired if learners are led by a blocked presentation to search for a verbalizable rule that does not in fact exist for such categories. Conversely, because interleaved presentation of exemplars makes the use of such a deliberate search strategy difficult, it

would presumably be less compatible for the learning of rule-based category learning and more compatible with the implicit learning system thought to underlie the learning of information-integration based categories.

It also seems reasonable to assume that learners might just intuitively adopt different types of learning strategies when faced with the task of learning naturalistic, perceptual-based categories versus conceptual, rule-based categories. From the learner's standpoint, given artificial or conceptual categories to learn—such as different types of statistical tests—it might just seem reasonable that the differences among such categories would likely be based on underlying rules, and thus a natural starting point in learning to discriminate among them would be to search for those rules. In contrast, given more naturalistic, perceptual-based categories to learn—such as artists' painting styles, bird families, or butterfly species—it might seem reasonable to learners just to assume that differences among such categories would not be defined by easily verbalizable rules and, thus, to realize that searching for such rules would not be an effective strategy for learning to discriminate among them. If we consider artists' painting styles to be an "information-integration" type of category (as suggested by Zulkiply & Burt, 2012), then the learning of these categories might benefit from interleaving, not because of high within- and between-category similarity, but as a result of the lack of clear category-defining rules. If so, then the learning of different artists' painting styles should always benefit from interleaved study, or at the very least, we may not find such a clear interaction between category similarity structure and presentation schedule in the learning of these types of perceptual-based categories.

In summary, we examined the effect of category similarity structure (i.e., the similarity relations within and between the categories of a to-be-learned set of categories) on presentation schedule efficacy for the learning of non-parametric statistical tests and artists' painting styles in Experiment 1 and Experiment 2, respectively. In accordance with prior studies, we expected to find an interleaving benefit in the learning of our overlapped set of to-be-learned categories (given their low-within and high-between similarity construction) and a blocking benefit in the learning of our distinct set of to-be-

learned categories (given their high-within and low-between similarity construction). In Experiments 3A and 3B, we further tested the hypothesis that noticing the commonalities within a category and differences between categories is critical to category learning by varying interleaved exemplars of conceptual and perceptual categories in one of two ways: exemplars from different categories were juxtaposed randomly, or exemplars from different categories were juxtaposed (i.e., blocked) by content. The latter (i.e., blocked-content, interleaved-category) sequence may be particularly effective at drawing attention to the features most relevant for discriminating between categories, thus enabling easier dismissal of non-diagnostic features (e.g., the content). On the other hand, in the former (i.e., random-content, interleaved-category) sequence, it may take longer to realize that the content is irrelevant for discriminating between categories because all categories share identical content. In Experiment 4, we revised the methodology of Experiment 2 to test whether blocking of the distinct set of perceptual categories would also facilitate better discrimination between the studied categories and new categories with identical content.

Experiment 1

In Experiment 1, we investigated whether category similarity structure would affect the relative efficacy of blocked and interleaved presentation schedules for the learning of conceptual, rule-based categories—specifically, different types of non-parametric statistical tests. Table 1 illustrates the different category structures in terms of similarity relations that were employed in Experiments 1 and 2. The top right cell illustrates the category similarity structure of the set of distinct categories and the bottom left cell illustrates the category similarity structure of the set of overlapped categories used in Experiments 1 and 2. The bottom right cell illustrates a set of categories added to the design of Experiment 2.

Table 1. *Category Similarity Structures among the Set of To-Be-Learned Categories in Experiment 1 (Distinct and Overlapped) and in Experiment 2 (Distinct, Overlapped, and Landscapes)*

Condition	Between-category similarity	Within-category similarity	Description
Distinct	Low	High	Each category has its own distinct content
Overlapped	High	Low	Every category has exemplars of every content
Landscapes	High	High	All artists' paintings are of landscapes (Exp 2 only)

Method

Participants and design. Ninety-one participants (age range: 19 - 70 years; mean age: 37.55 years; median age = 34; 50 females) were recruited from Amazon Mechanical Turk (MTurk) and paid $1.00 for their participation. We manipulated presentation schedule (blocked vs. interleaved) and category similarity structure (distinct vs. overlapped) in a 2x2 factorial design with all variables manipulated between subjects.

Materials. All participants studied three word-problem exemplars illustrating each of three non-parametric statistical concepts: Chi-squared test, Kruskal-Wallis test, and Wilcoxon signed-rank test. Then, in the following classification test phase, participants were asked to identify which of these three types of statistical concepts was represented by six randomly selected and never-before-studied word-problems, with the specific word-problems used in the test phase kept constant across conditions. The study exemplars and test exemplars were all between 88–95 words in length and did not include computational solutions, as their primary purpose was to demonstrate the type of scenario or design that would be appropriately analyzed by the different types of statistical tests or concepts.

For the set of to-be-learned categories with a distinct category similarity structure, the storyline presented in each of the studied exemplars for a given statistical concept was different. Specifically, the storyline for the word problems appropriate for a Chi-Square test were about schooling and education; those appropriate for a Kruskal-Wallis test were about fashionable apparel; and those appropriate for a Wilcoxon signed-rank test were about fruits and vegetables. Thus, for the set of distinct to-be-learned categories, the category similarity structure was one in which within-category similarity was high and the between-category similarity was low. For the set of to-be-learned categories with overlapped content, one exemplar of each of the content storylines (schooling and education, fashionable apparel, fruits and vegetables) was studied for each concept. In other words, the exemplars for a given statistical concept consisted of three different storylines, and each storyline appeared in the three exemplars of each statistical concept. Thus, for the set of overlapped to-be-learned categories, the category similarity structure was one in which within-category similarity was low and the between-category similarity was high.

The problems appearing on the final classification test, which were kept identical across conditions, were taken from various statistics textbooks, were different in content to those of the study exemplars, and were also different from each other.

Procedure. Participants were instructed that their task was to study word-problem exemplars illustrating three different statistical concepts, such that they would be able to identify (from a list of names) the appropriate statistical test illustrated by new, never-before-studied word problems on a later final classification test. In the study phase, exemplars were presented sequentially on a computer screen for 25 s each. When the presentation of exemplars was blocked, all the exemplars of a given statistical concept were presented consecutively, with the order of the concepts and the order of the specific exemplars of a given concept randomized for each individual. When the presentation of exemplars was interleaved, the nine exemplars were organized in blocks of three exemplars, with each block containing one randomly selected exemplar from each concept.

The final classification test followed a 60-s crossword puzzle distractor task. In the final test, participants were shown six new word-problems of each concept and asked to select, from a list containing the names of the three studied statistical concepts, the one most appropriate for the presented problem. The test problems were presented on a computer screen in a random order with the exception that no two problems representing the same concept were presented consecutively. As soon as participants made their choice of the concept most appropriate for a given problem by clicking on a name from the list, the next problem to be classified appeared on the screen until all test problems had been presented. The test was self-paced and included no feedback.

Finally, all participants were debriefed regarding the two types of presentation schedules (blocked vs. interleaved) used in Experiment 1. The two schedules were described, participants were reminded of the presentation schedule that they had experienced, and then they were asked which schedule (interleaved or blocked) they believed would lead to the best learning of the statistical concepts.

Results and Discussion

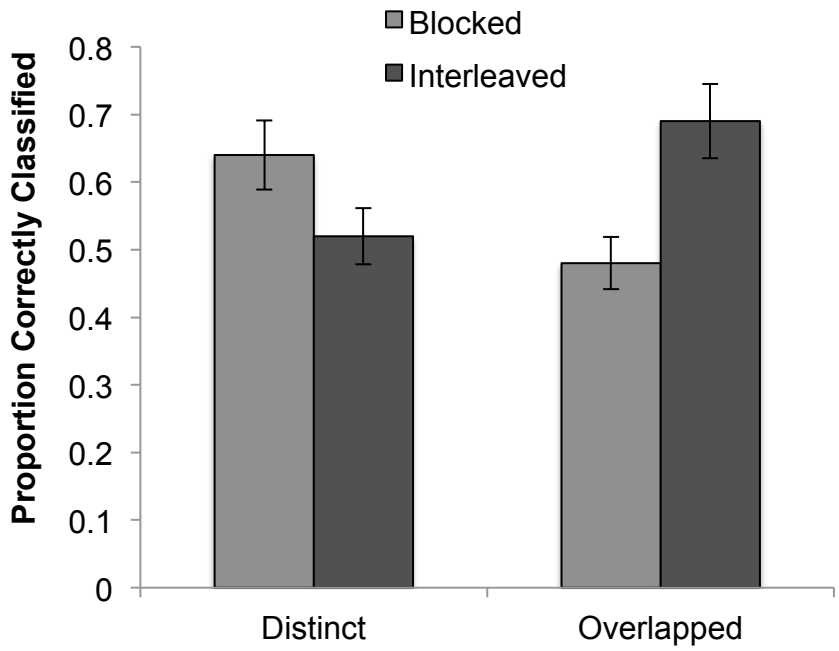

Figure 1. The proportion of new problems correctly classified on the final test for the two types of category similarity structures (distinct vs. overlapped) and the different

presentation schedules (blocked vs. interleaved) in Experiment 1. Error bars represent standard error of the mean.

Classification performance. Performance obtained on the classification test as a function of category similarity structure and presentation schedule is illustrated in Figure 1. A 2 (presentation schedule: blocked vs. interleaved) x 2 (category similarity structure: overlapped vs. distinct) between-subjects analysis of variance (ANOVA) performed on the data revealed no significant main effect of presentation schedule, $F(1, 87) = .1.34$, $MSE = .04$, $p = .250$, $\eta_\rho^2 = .02$, although average performance obtained with an interleaved presentation schedule ($M = .61$, $SD = .22$) was numerically better than that obtained with a blocked presentation schedule ($M = .56$, $SD = .22$). Additionally, no significant main effect of category similarity structure was observed, $F(1, 87) = .02$, $p = .88$, $MSE = .04$, $\eta_\rho^2 = .00$. That is, overall performance did not differ for the overlapped ($M = .59$, $SD = .24$) and distinct ($M = .58$, $SD = .21$) sets of to-be-learned categories. Critically, however, a significant interaction between category similarity structure and type of presentation schedule did emerge, $F(1, 87) = 14.96$, $p < .0001$, $MSE = .04$, $\eta_\rho^2 = .15$. As predicted, pairwise comparisons revealed that when the content of exemplars overlapped for each concept (i.e., when within-category similarity was low and between-category similarity was high for the set of categories to be learned), participants' ability to learn the different concepts or categories was better given an interleaved ($M = .70$, $SD = .23$) rather than a blocked ($M = .48$, $SD = .19$) presentation schedule, $t(44) = 3.50$, $p = .001$, $d = 1.03$. Conversely, when the content of the exemplars for each concept was distinct (i.e., when within-category similarity was high and between-category similarity was low for the set of categories to be learned), participants given a blocked presentation schedule ($M = .64$, $SD = .22$) outperformed those given an interleaved schedule ($M = .52$, $SD = .18$), $t(43) = 1.95$, $p = .058$, $d = .59$.

Metacognitive judgments. Overall, a majority (64%) of participants judged that a blocked schedule of learning was or would have been better for their own learning than an interleaved schedule, and only 36% reported thinking that interleaving was or would be better for learning. The pattern of metacognitive judgments did not differ by type of

category similarity structure (i.e., overlapped vs. distinct content), $\chi^2(4) = .33$, $p = .66$, or by type of presentation schedule (blocked vs. interleaved), $\chi^2(2) = .09$, $p = .82$

Experiment 2

Experiment 2 differed from Experiment 1 in that the to-be-learned categories were perceptual-based—specifically, different artistic styles—rather than the rule-based, statistical categories used in Experiment 1. Experiment 2 contained a conceptual replication of Experiment 1 in that two sets of the to-be-learned categories had the same similarity structures as the distinct and overlapped categories. Experiment 2, however, also included a third set of to-be-learned categories for which both the within-category and between-category similarity was high —in essence, then, the structure for the sets of to-be-learned categories in most of the studies from the literature that we previously described (e.g., Kang & Pashler, 2011; Kornell & Bjork, 2008). As all the exemplars in this set of to-be-learned categories were paintings of landscapes, we henceforth refer to the set of to-be-learned categories with this type of similarity structure (i.e., both high within- and high between-category similarity) as the *landscapes* category.

Method

Participants and design. Two hundred and eighty-four participants were recruited from Amazon Mechanical Turk and paid $0.40 for their participation. Seven MTurk participants were eliminated from analyses as they indicated having prior experience with the study materials, resulting in a remaining total of 276 participants (age range: 18-50 years; mean age: 30.2 years; median age = 29; 139 females). As in Experiment 1, two levels of presentation schedule were used (*blocked* vs. *interleaved*). In addition to assessing how presentation schedule interacted with the two types of category similarity structures examined in Experiment 1 (i.e., distinct vs. overlapped), Experiment 2 also assessed how presentation schedule might interact with the learning of a set of categories with a third type of category similarity structure—namely, one in which both the within-category and the between-category similarity is high. The final row of Table 1 illustrates this third type of high within and between category similarity of the to-be-learned categories in this set, which, as previously mentioned, we refer to as the

landscapes category set. Experiment 2 thus employed a 2 (presentation schedule: blocked vs. interleaved) x 3 (category similarity structure: overlapped vs. distinct vs. landscapes) factorial design with all variables manipulated between-subjects.

Materials. All the paintings used in Experiment 2 as exemplars for the different to-be-learned categories of artistic styles (as well as those used in Experiments 3B and 4) were drawn from www.dailypainters.com. Each participant studied eight paintings (sized to be as close to 500x400 pixels as possible) by each of four artists (Toni Grote, Jamie Grossman, Julie Ford Oliver, and Gerald Schwartz), and they were tested on four never-before-studied landscape paintings from each of these artists in the following classification test. The four never-before-seen landscape paintings by each artist that appeared in the classification test were chosen randomly and held constant across all conditions.

For participants assigned to the landscapes category set, all the exemplars presented during the study phase were paintings of landscapes so that there was both high within- and between-category similarity for the to-be-learned categories in this set (analogous to the materials used in Kornell & Bjork, 2008). For participants assigned to receive exemplars from the set of distinct categories during the study phase, the four paintings presented as exemplars of a given artist's style all contained the same content, and this content was different from that of all the other artist's paintings. Specifically, Grote's paintings were always of buildings, Grossman's paintings were always of flowers, Oliver's paintings were always of pots and containers, and Schwartz's paintings were always of food. Thus, for these participants, there was high within-category similarity and low between-category similarity among the set of categories that they were to learn. Finally, for participants assigned to receive exemplars from the set of overlapped categories during the study phase, two paintings of each content type (buildings, flowers, pots and containers, and food) were studied for each artist. Thus, for these participants, there was low within-category similarity and high between-category similarity among the set of categories that they were to learn. An illustrative sample of the stimuli used in these three types of category similarity structures (distinct vs.

overlapped vs. landscapes) is presented in Figure 2 (specifically, half of the paintings for half of the artists).

Figure 2. Illustrative exemplars from the three category similarity structures—landscapes (high within- and between-category similarity), distinct (high within- and low between-

category similarity; each artist paints a different content), and overlapped (low within- and high between-category similarity; each artist paints each content)—used in Experiment 2.

Procedure. Participants were told that their task was to learn to recognize the painting styles of four different artists, such that they would be able to identify (from a list of their names) the artist responsible for new, never-before-studied paintings by the artists on a final classification test. In the study phase, images were presented sequentially on a computer screen for 3 s each with the name of the artist appearing below each painting. When the paintings were presented in a blocked schedule, all the paintings by a given artist were presented consecutively, with the presentation order of the four artists and their specific paintings randomized for each participant. When the paintings were presented in an interleaved schedule, the 32 to-be-studied paintings were organized into four blocks of eight paintings, with two randomly selected paintings by each artist appearing in each block. From the perspective of the participant, however, the study phase appeared as just one long sequence of 32 randomly arranged paintings.

Following the study phase and a 45-s distractor phase in which participants played Tetris, the final classification test began. For all participants, regardless of the content of the exemplars they were shown during the study phase, the new paintings by each artist presented in the final classification test were of landscapes. The test images were presented in two randomized blocks of four paintings, with each block containing one painting by each artist. As soon as participants made their choice of the artist responsible for a given painting by clicking on a name from the list, the next painting to be classified appeared on the screen until all test paintings had been presented. The test was self-paced and included no feedback.

When the exemplars studied were from the set of distinct categories (i.e., the set of categories with high within-category similarity and low between-category similarity), participants were also asked some additional questions about the different artist's styles following completion of the classification test. Making their choices simultaneously on a single webpage, they were asked to select from a list (flowers, food, buildings, or pots

and containers) the content of each artist's studied paintings. They were then asked, "Did you notice the mapping between artist and content of the paintings (e.g., Grossman = flower, Schwartz = food, etc.)? If you noticed it, did this mapping help, hurt, or make no difference to your ability to learn each artist's style?" and they indicated their answers by selecting from a given list of responses (i.e., did not notice; helped; hindered; made no difference).

Finally, all participants were debriefed regarding the two presentation schedules used in the study. The two schedules were described, participants were reminded of the presentation schedule that they had experienced, and then they were asked which schedule they believed would lead to the best learning of the different artistic painting styles (interleaved, blocked, or no difference).

Results and Discussion

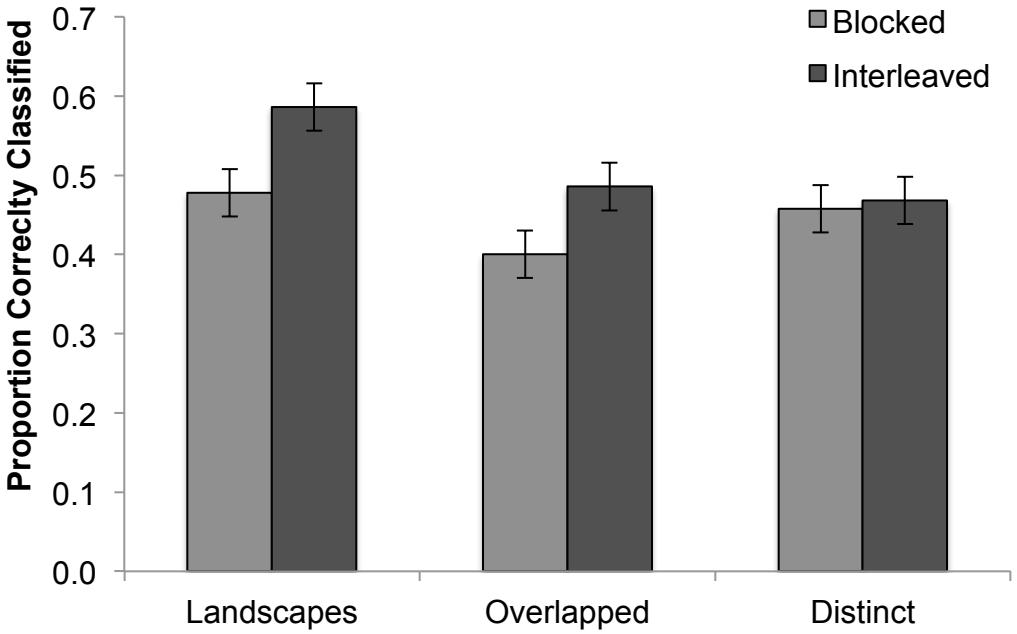

Figure 3. The proportion of new paintings correctly classified on the final test for the three types of category similarity structures (landscapes vs. overlapped vs. distinct) and the different presentation schedules (blocked vs. interleaved) in Experiment 2. Error bars represent standard error of the mean.

Classification performance. The average proportion of new paintings correctly classified on the final test as a function of category similarity structure and presentation schedule is illustrated in Figure 3. A 2 (presentation schedule: blocked vs. interleaved) x 3 (category similarity structure: landscapes vs. distinct content vs. overlapped between-subjects ANOVA performed on the data revealed both a significant main effect of presentation schedule, $F(1, 270) = 7.22$, $MSE = .04$, $p = .01$, $\eta_\rho^2 = .03$ and a significant main effect of category similarity structure , $F(1, 270) = 4.27$, $MSE = .04$, $p = .02$, $\eta_\rho^2 = .03$. The main effect of presentation schedule replicated the findings of prior studies using these types of materials: Average final test performance obtained with an interleaved presentation schedule ($M = .51$, $SD = .19$) was better than that obtained with a blocked presentation schedule ($M = .45$, $SD = .18$).

Post-test comparisons were used to examine the effect of category similarity structure: Overall, participants receiving the landscape category set performed better on the final classification test ($M = .54$, $SD = .23$) than did the participants studying exemplars from the distinct category set ($M = .46$, $SD = .20$) or the participants studying exemplars from the overlapped category set ($M = .44$, $SD = .21$), $t(178) = 2.30$, $p = .02$, Cohen's $d = .34$ and $t(176) = 2.88$, $p < .01$, Cohen's $d = .43$, respectively. Performance of the participants who had studied exemplars from the overlapped set of categories versus exemplars from the distinct set of categories did not differ significantly from each other, $t(192) = .69$, $p = .49$, Cohen's $d = .08$.

The interaction between presentation schedule and category similarity structure was not found to be significant, $F(1, 270) = 1.37$, $MSE = .04$, $p = .26$, $\eta_\rho^2 = .01$. Given our hypotheses, however, we conducted three planned comparisons, examining the effect of schedule within each type of category similarity structure. In the landscapes category structure (high within- and high between-category similarity among the to-be-learned categories), final classification performance was significantly better after interleaved study ($M = .59$, $SD = .22$) as compared to after blocked study ($M = .48$, $SD = .24$), $t(80) = 2.17$, $p = .03$, Cohen's $d = .48$, consistent with results of prior studies (e.g., Kang & Pashler, 2012; Kornell & Bjork, 2008; Kornell, Castel, Eich, & Bjork, 2010). More

critically, the effect of schedule differed between the overlapped and distinct category structures. For the learning of the set of categories with an overlapped structure, there was, as predicted, a significant benefit of an interleaving presentation schedule (M = .49, SD = .21) over a blocking presentation schedule (M = .40, SD = .20), $t(94) = 2.06$, p = .04, Cohen's d = .42. For the learning of the set of categories with a distinct structure, however, contrary to the prediction that a blocking presentation schedule should be more effective for learning than an interleaving presentation schedule, classification accuracy did not differ between the blocked (M = .46, SD = .20) and interleaved (M = .47, SD = .20) presentation schedules, $t(96) = .26$, $p > .05$, Cohen's d = .05.

Content-mapping performance for the distinct category structure. Participants asked to learn the set of categories with the distinct category structure were also asked to identify the content of the paintings by each artist (selecting their response from a list of four options). The ability of participants to identify the content associated with each artist did not differ with respect to whether they had studied under a blocked presentation schedule (M = .78, SD = .30) or an interleaved presentation schedule (M = .78, SD = .31), $t(96) = .94$, p = .94. Furthermore, when asked whether they thought that the mapping had affected their learning, their responses were not dependent on the presentation schedule experienced, $\chi^2(3) = 1.50$, p = .68. Overall, 37% of the participants reported thinking that the mapping had hindered their learning of the different artists' styles, 35% reported thinking that it had helped them learn the different artists' styles, 17% reported thinking that it made no difference, and 11% reported that they had not noticed the mappings.

Metacognitive judgments. Overall, a majority (66%) of the participants reported thinking that a blocked schedule of learning was best for their own learning, and only 20% reported thinking that interleaving was better for learning. Moreover, the pattern of metacognitive judgments did not differ by category similarity structure, $\chi^2(4) = 6.38$, p = .17, nor by experienced presentation schedule, $\chi^2(2) = .75$, p = .69. Interestingly, however, the pattern of judgments seemed to differ marginally between those participants who had experienced a blocked versus an interleaved schedule of presentation while

learning the set of categories with an overlapped category structure, $\chi^2(2) = 4.90$, $p = .09$: For those learning under a blocked presentation schedule, 60% reported that blocking would be most effective, but for those who had learned under an interleaved presentation schedule, this figure rose to 81% of participants. In sum, participants were insensitive to the relative efficacy of the presentation schedules, regardless of category structure.

Experiments 3A and 3B

In Experiments 1 and 2, for the participants learning the set of overlapped categories (i.e., low within-category and high between-category similarity) and receiving an interleaved presentation sequence of the exemplars during study, the exemplars representing the different statistical concepts and artists' styles were presented in a more or less randomized order. That is, although organized in blocks containing only one (Experiment 1) or two representations (Experiment 2) of each statistical concept or artist's style, respectively, with those exemplars randomly ordered within each block, the study phase would have undoubtedly appeared as just one long sequence of randomly arranged problems or paintings from the participant's perspective.

We could, however, have organized the sequence of exemplars so that—while keeping the presentation of exemplars from the different categories interleaved—the content of the exemplars was either massed or randomized. For example, in the case of learning artists' painting styles, although the artists are interleaved, the content of those paintings can be blocked—learners could study all the paintings containing buildings, before moving on to the paintings containing flowers, and so on (i.e., massed content). Or, they could study all the paintings intermixed (i.e., shuffled content). If, as hypothesized, the learning of a set of categories with this type of overlapping similarity structure (i.e., low within-category and high between-category similarity) depends critically on the learner gaining an understanding of how the categories differ, massing the content should very quickly emphasize to the learner the irrelevance of the content, reduce the variance that arises naturally from differing content, and therefore focus attention more quickly on the key category differences.

We assessed this possibility in Experiment 3A, using the non-parametric statistical concepts as the to-be-learned categories, by presenting exemplars of the different concepts with similar storylines together (i.e., massed content). By doing so, learners should be able to eliminate the "noise" introduced when both content and conceptual features differ between consecutively presented exemplars, freeing them to focus on the critical, structural differences between the consecutively presented different concepts, and thus their induction of the different concepts should be improved. In Experiment 3B, we similarly assessed whether a benefit on induction of artists' painting styles would obtain if we massed the content of the paintings while still interleaving the paintings of the different artists. We predicted that by presenting, for instance, all the flower paintings by all of the artists consecutively, learners would be able to eliminate the "noise" introduced when both content and style differed between consecutive paintings and, again, be freed to focus on the critical stylistic differences between the artists.

Method

Participants and design. A total of 45 participants (27 females; age range = 20-66; mean age = 37.91; median age = 34) were recruited from Amazon Mechanical Turk and paid $1.00 for their participation for Experiment 3A, and 48 participants (22 females; age range = 18-66; mean age = 35.25; median age = 32.50) were recruited from Amazon Mechanical Turk and paid $0.40 for their participation for Experiment 3B. Five participants were eliminated from analyses in Experiment 3B for indicating having participated in a similar study using these materials previously. Whereas all exemplars were presented for study in an interleaved manner, whether the content of the exemplars was massed or shuffled during study was manipulated between subjects.

Materials. For Experiment 3A, the study and test materials were the same as those used for the set of overlapped categories in Experiment 1; for Experiment 3B, the study and test materials were the same as those used in the set of overlapped categories in Experiment 2.

Procedure. For participants assigned to the shuffled content condition of Experiment 3A, the procedure was identical to that for participants studying the set of

overlapped categories with an interleaved presentation in Experiment 1. For participants assigned to the shuffled content condition of Experiment 3B, the procedure was identical to that for participants studying the set of overlapped categories with an interleaved presentation in Experiment 2. For participants assigned to the massed content condition of Experiment 3A, exemplars were presented in blocks of three with one exemplar for each statistical concept and with all the exemplars in that block featuring the same storyline—that is, all were about schooling and education, or about fashionable apparel, or about fruits and vegetables. For participants assigned to the massed content condition of Experiment 3B, exemplars were presented in blocks of eight paintings with two paintings by each artists in each block, which were randomly ordered except that paintings by the same artist could not be consecutively presented, and all eight paintings featured the same content—that is, all were paintings of flowers, or of food, or of buildings, or of pots and containers. Described another way, participants in the massed content condition saw all the images of one content type (say, all the paintings of flowers) before seeing those of another content type (say, all the paintings of food). The order in which the different content types were presented was randomized across participants in both Experiments 3A and 3B.

The final classification tests in Experiment 3A and 3B were the same as those used in Experiment 1 and 2, respectively. The test was self-paced, included no feedback, and all test exemplars were presented one at a time. In Experiment 3A, the final classification test followed a 60-s crossword puzzle distractor task. Participants were shown six new word-problems of each concept and asked to select, from a list containing the names of the three studied statistical concepts, the one most appropriate for the presented problem. No two problems representing the same concept were presented consecutively. In Experiment 3B, the classification test began following the study phase and a 45-s Tetris distractor task. The test images were new landscape paintings by each artist, and were presented in two randomized blocks of four paintings, with each block containing one painting by each artist. Participants made their choice of the artist

responsible for a given painting by clicking on a name from the list of names presented below each painting.

Results and Discussion

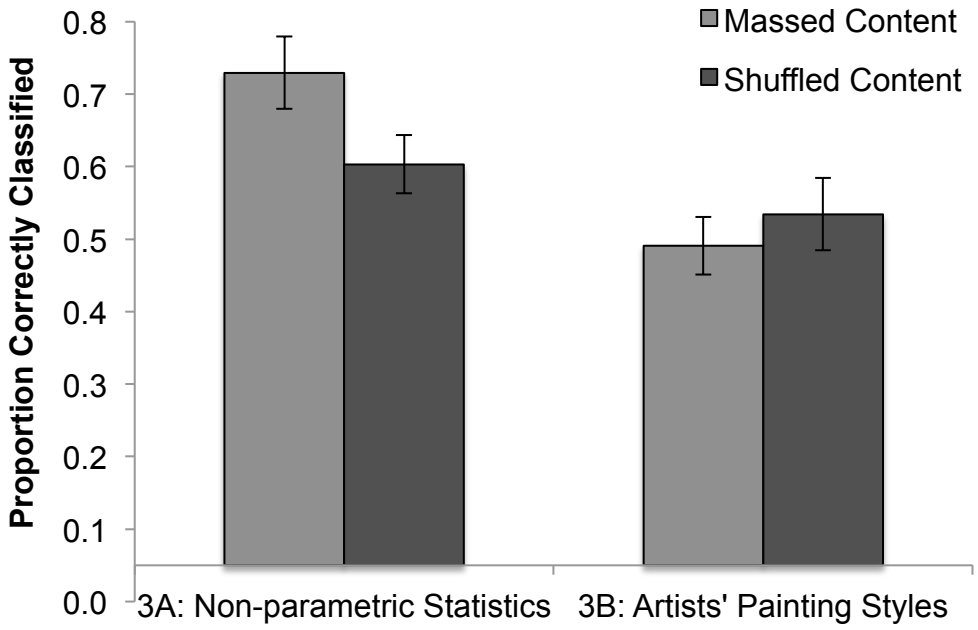

Figure 4. Classification test performance for each type of content sequencing (massed vs. shuffled) in Experiments 3A and 3B. Error bars represent standard error of the mean.

Correct performance obtained on the classification test as a function of content sequencing for the two types of to-be-learned categories (non-parametrical statistical tests in Experiment 3A or artists' painting styles in Experiment 3B) is illustrated in Figure 4. As indicated there, learning of the non-parametric statistical categories (Experiment 3A) was better when the content (i.e., storyline) of the different to-be-learned concepts was presented in a massed (M = .68, SD = .23) rather than a shuffled manner (M = .55, SD = .17), $t(43) = 2.08$, $p = .044$, $d = .64$. In Experiment 3B, however, the learning of the different artists' styles did not differ significantly as a function of whether the content of the paintings by the different artists was shuffled (M = .48, SD = .23) or massed together (M = .44, SD = .18) in the study phase, $t(41) = .75$, $p = .46$, Cohen's d = .21. If anything, the pattern of results obtained is opposite to that found in Experiment 3A as well to our predictions.

Experiment 4

For the set of to-be-learned distinct categories in Experiment 2 (i.e., the set for which within-category similarity was high and between-category similarity was low), we had predicted that a blocked presentation schedule would lead to better induction of the different artists' painting styles than would an interleaved presentation schedule. Contrary to this prediction, classification performance did not differ as a function of the study schedule. The paintings used in the test phase of Experiment 2, however, were all of landscapes, not paintings having the same content as those presented during the study phase, which might have obscured any benefit that a blocked schedule, as compared to an interleaving one, might have bestowed on the learning of this set of categories. In Experiment 4, we examined this possibility by assessing the learning of this same set of categories (i.e., distinct) under a blocked versus an interleaved presentation schedule, but we switched to presenting paintings in the classification test whose content was consistent with the artist-content mappings of the paintings shown to participants during study.

As in Experiment 2, we expected that the participants, whether given an interleaved or a blocked schedule during study, would learn the artist-content mappings present in the set of paintings they studied. We speculated, however, that a blocked schedule, might lead to better learning of the specific styles of the different artists beyond just learning the different contents painted by the different artists. We reasoned that if an interleaved study schedule focused learners on the processing of differences between the artists, then simply realizing the artist-content mapping present in the set of studied paintings would be sufficient to distinguish between the artists. In a blocked study schedule, however, perhaps the learners' attention would also be drawn to the commonalities within a given artist's paintings—that is, not only to the specific content of a given artist's paintings (which they should notice faster than would the participants studying the paintings in an interleaved manner) but also to the artist's style of painting that content. We thus conjectured that participants studying under a blocked schedule would become better able than those studying under an interleaved schedule to distinguish among new paintings of the same content done by the studied artists versus

those done by new, unstudied artists —that is, to distinguish between flowers painted by Grossman, for example, and flowers painted by some new artist. More specifically, we might expect the interleaved schedule to produce more false alarms to lures.

Method

Participants. Eighty-nine participants (44 females; age range = 18-64; mean age = 35.03; median age = 32) were recruited from Amazon Mechanical Turk and paid $0.45 for their participation.

Materials. The set of categories to be learned in Experiment 4 was the same as the set of distinct categories (i.e., high within and low between category similarity) used in Experiment 2. Thus, the participants of Experiment 4 saw the same eight paintings by each of the four artists (Grote, Grossman, Oliver and Schwartz) as had been seen by the participants of Experiment 2, with each artist painting a different content (i.e., buildings by Grote; flowers by Grossman; pots/containers by Oliver; food by Schwartz).

The nature of the test phase given to the Experiment-4 participants, however, was entirely different from that of the test phase for the Experiment-2 participants. Specifically, rather than being tested on landscapes painted by the studied artists, the content of each painting appearing in the test phase for the Experiment-4 participants was comprised of one of the four contents of the studied paintings (i.e., flowers, pots/containers, food, or buildings). Thus, for example, all of the new paintings of Grossman appearing in the test phase were new paintings of flowers—that is, they were of the same content as all of the studied paintings by Grossman, rather than being landscapes by Grossman.

Additionally, because the question being assessed in Experiment 4 was whether participants studying the paintings of each artist in a blocked manner (as opposed to an interleaved manner) would be better at distinguishing the individual styles of the studied artists from those of non-studied artists painting the same content, paintings of non-studied artists (incorrect lures) were also presented during the test phase. And, in contrast to the task of the Experiment-2 participants, which was to classify the styles of the studied artists, the task of the Experiment-4 participants was to judge for each presented painting

whether it was one that had been painted by a studied artist or one painted by some non-studied artist.

The lures (or paintings by non-studied artists) presented in the classification test were selected to be very similar in content to their yoked artists and reasonably similar in style. To illustrate, Figure 5 shows a side-by-side comparison of studied paintings by Schwartz and the lures for Schwartz's paintings that were presented on the final classification test.

Procedure. The procedure of the study phase in Experiment 4 was identical to that for the participants learning the set of distinct categories in Experiment 2. For the test phase, the procedure was different—the final test consisted of recognition and classification questions, was self-paced and included no feedback. The Experiment-4 participants were informed that they would be shown paintings that they had not studied, with half of these new paintings being ones by one of the four studied artists (Grossman, Grote, Oliver, or Schwartz), and the other half being ones by new artists—that is, painted by artists that were not one of the four studied artists. For recognition questions, participants were presented with the test paintings one at a time and with two buttons, one labeled "Studied Artist" and one labeled "Not Studied Artist (maybe New Artist)" that appeared on the computer screen beneath each painting. If participants thought the painting was one that had been painted by a studied artist, they were to select the "Studied Artist" button; if they thought it was not a painting by one of the studied artists, they were to select the "New Artist" button. For any test painting indicated to be by a studied artist, the participants were prompted with a classification question—they were asked to select which of the studied artists they believed was responsible for that painting from a provided list of their names. The classification question appeared only after participants selected and submitted the "Studied Artist" response. If instead, the participants indicated thinking that the presented painting was by a new artist, they were not asked to make any further responses, and the next test painting was immediately presented. A block-randomization procedure was used to determine the presentation order of the paintings in the test phase. Each "block" consisted of eight paintings (one by each of the four studied

artists plus one lure for each studied artist), and the presentation order of the eight paintings within each block was randomized. A single sequencing of test paintings was generated consistent with this block randomization procedure, and then this sequence was kept constant across all participants.

After all the test paintings had been presented, the Experiment-4 participants were asked the same post-test questions as were asked of participants studying the set of distinct categories in Experiment 2.

Schwartz's Studied Paintings	Unstudied Artist Test Paintings (Lures for Schwartz)
	
	
	
	

Figure 5. A side-by-side comparison of the studied paintings and final-test lures for one of the four artists (Schwartz) used in Experiment 4.

Results and Discussion

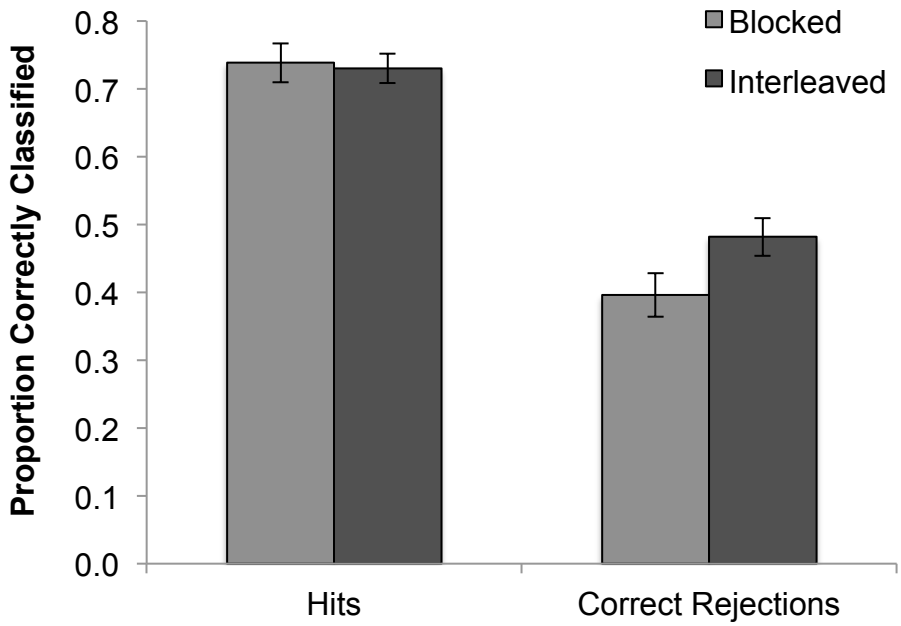

Figure 6. Participants' ability to distinguish between paintings by studied artists versus new artists, as measured by hit and correct-rejection rates obtained on the recognition component of the final test for Experiment 4.

Recognition performance. Participants' performance on the recognition part of the final test, measured in terms of their hit and correct-rejection rates, is shown in Figure 6. As indicated there, average hit rates did not significantly differ for participants who had studied in a blocked manner (M = .74, SD = .20) versus for those who had studied in an interleaved manner (M = .73, SD = .14), $t(87)$ = .22, p = .83, Cohen's d = .05. The correct rejection rates, however, were marginally better for participants who had studied in an interleaved manner (M = .48, SD = .18) versus a blocked manner (M = .40, SD = .22), $t(87)$ = 1.98, p = .05, Cohen's d = .43. This suggested interaction between type of presentation schedule and type of judgment, however, did not reach significance, $F(1,87)$ = 1.78, MSE = .06, p = .19, η_p^2 = .02.

Classification performance. There was a significant main effect of painting type (new painting by a studied artist vs. by a new artist): Participants were better able to identify the appropriate name for paintings by studied artists (M = .83, SD = .24) than for paintings by new artists incorrectly identified as being by studied artists (M = .73, SD = .24), $F(1, 87) = 29.39$, $p < .001$, $\eta_p^2 = .25$.

On test trials in which participants selected "Studied Artist" for a presented painting, they were subsequently prompted to select the name of the studied artist they thought was responsible for that specific painting from a list of names of the studied artists. For this classification component of the final test, there were no significant differences in the ability to select the appropriate name (i.e., the correct artist's name for the paintings judged to be by studied artists and actually were by studied artists, or the incorrect artist's name for lure paintings judged to be by studied artists but actually were by some unfamiliar artists). For paintings correctly identified as being by a studied artist, participants who had studied in an interleaved manner were numerically, but not significantly, better at identifying the correct artist's name (M = .85, SD = .24) than were the participants who had studied in a blocked manner (M = .80, SD = .26), $t(87) = .97$, p = .33, Cohen's d = .21. For paintings incorrectly identified as being by a studied artist, participants who had studied in an interleaved manner (M = .72, SD = .30) were numerically less likely to select the name of the studied artist for whom that painting had been selected as a lure than were the participants who had studied in a blocked manner (M = .74, SD = .24), $t(87) = .35$, p = .73, Cohen's d = .07.

Content-mapping performance. The ability of participants to identify the type of content associated with the paintings of each studied artist (flowers, pots/containers, food, or buildings) did not differ with respect to whether they had studied in a blocked manner (M = .81, SD = .29) or in an interleaved manner (M = .84, SD = .30), $t(87) = .54$, p = .59, Cohen's d = .10. Similarly, when asked whether they thought that the mapping had affected their learning, their responses were not dependent on the presentation schedule they had experienced, $\chi^2(3) = 5.10$, p = .16. Overall, 44% of the participants reported that mapping had hindered their learning of the artists' styles, 33% reported that

it had helped them to learn the artists' styles, 17% reported that it had made no difference, and 18% reported that they did not notice the mappings. This pattern of responses is similar to the one observed in Experiment 2.

Metacognitive judgments. With respect to the questions regarding which type of presentation schedule they thought would be more effective for their learning (given the choices interleaved, blocked, or no difference), the majority (60%) of participants reported post-test that they thought a blocked schedule of presentation would be more beneficial for their inductive learning than an interleaved schedule would be, and only 24% of the participants reported thinking that an interleaved presentation schedule would be more effective. Again, participants' judgments were not dependent on experienced schedule, $\chi^2(2) = 3.69$, $p = .16$. It is interesting to note, however, that participants who experienced the interleaved presentation schedule were less likely to believe that blocking would be the more effective schedule for learning than were the participants who experienced the blocked presentation schedule (49% versus 69%, respectively) and, furthermore, they were more likely to believe that interleaving would be the more effective schedule for learning (29% vs. 19%, respectively).

General Discussion

The goal of the present study was to examine the possible interaction of two factors as they affect the learning of conceptual and perceptual categories: category similarity structure (i.e., the within- and between-category similarity relations among the set of categories to be learned) and presentation schedule (interleaved and blocked). We predicted that blocked study would promote the noticing of within-category commonalities and interleaved study would promote the noticing of between-category, and that the optimal schedule for learning would depend on the category similarity structure. Consistent with these predictions, we demonstrate that the efficacy of a given schedule does indeed seem to depend on the similarity of the category structures. Moreover, the specific effects also appear to partly vary depending on whether the categories are rule-based or perceptual-based.

With the conceptual, rule-based categories (i.e., non-parametric statistics concepts), the Experiment 1 results were consistent with our original predictions: an interleaving schedule led to better induction when content of the category sets overlapped (i.e., there was high between-category similarity and low within-category similarity, where noticing differences between categories was critical); a blocked schedule led to better induction when content of each category set was distinct (i.e., there was low between-category similarity and high within-category similarity, where noticing the similarities within-categories was critical). In line with the notion that the benefit of category learning depends on the salient characteristics that a schedule emphasizes, Experiment 3A showed that when stimuli were drawn from overlapping categories, an interleaved schedule could be made more or less effective by manipulating the specific juxtapositions of the content—massing the content allowed participants to more easily disregard the irrelevant features and focus instead on the critical differences between categories compared to when the content was shuffled.

With the perceptual categories (i.e., the artistic painting styles), the Experiment 2 results were less consistent with regard to our original predictions: an interleaved schedule led to better induction when content of the category sets were overlapped, as predicted; however, a blocked schedule did not lead to better induction when content of each category set was distinct, even though the interleaving advantage was eliminated, In Experiment 4, with the content of each artist's paintings distinct from the content of the other artists, we created a situation in which the blocked schedule should produce a greater advantage over the interleaved schedule. We did this by testing learners' on their ability to distinguish between paintings by studied artists and paintings by unstudied artists with similar content (i.e., lures). We predicted that interleaved schedules would draw learners' attention to key differences (i.e., the content) between the artists, whereas blocked schedules would draw their attention to the similarities between an artist's paintings (i.e., their painting style beyond just the superficial content details). Therefore, the blocked schedule should lead to better discrimination between the studied artists' paintings and the content-lures. No such difference in classification test performance,

however, was observed. Similarly, in Experiment 3B, where the overlapped artists' paintings were presented interleaved, classification performance was similar regardless of whether the content of exemplars was massed or shuffled, despite our initial predictions that massing the content should reduce superficial noise and highlight critical differences between artists (as was found in Experiment 3A).

Taken together, the results from the current study contribute to the growing body of evidence to suggest that category similarity structure modulates the interleaving effect. The benefit of presentation schedules depends on whether the emphasis is to discriminate between exemplars of different categories or to encode commonalities of exemplars within categories. Moreover, the results suggest that the nature of the categories themselves may drive induction differently. This latter implication highlights the importance of considering the nature of stimuli used in experiments. In the current study, we cannot isolate what precisely it is about the nature of the categories that leads to differential inductive processes, but one potential candidate is the dual-process framework of category induction (Ashby & Maddox, 2011), which differentiates between rule-based learning and information-integration learning. Rule-based learning is thought to be an explicit process in which learning is best supported by a strategy of hypothesis-testing and rule discovery. In the prior studies using artificial stimuli (e.g., Carvalho & Goldstone, 2014; Zulkiply & Burt, 2012) and in our conceptual-based stimuli (i.e., non-parametric statistical tests), the categories are well defined by specific rules that learners must discover.

With these well-defined, rule-based categories, it appears that category similarity structure mediates schedule effects, such that interleaving outperforms blocking when between-category similarity is high, and blocking outperforms interleaving when between-category similarity is low. On the other hand, information-integration learning is thought to be an implicit process and facilitates learning in cases where the rules are not verbalizable—our artists' painting styles categories may fall more in line with this type of learning. Category similarity structure may not mediate schedule effects to the same extent for information-integration category learning, in part, potentially because blocking

encourages explicit hypothesis-testing (which is not optimal for implicit, information-integration learning) while interleaving makes such explicit hypothesis-testing difficult.

Further studies are warranted to replicate and more closely investigate the mediating effects of category similarity structure (i.e., within- and between-category similarities) and category type (i.e., rule-based vs. information-integration) on optimal presentation schedules, and how these variables may interact with different types of categories. So far, the experimental designs in previous studies (e.g., Carvalho & Goldstone, 2014; Zulkiply & Burt, 2012) and in the current study vary irrelevant features (e.g., content of statistical concepts' descriptions and artists' paintings) to examine possible interactions between category similarity structure and presentation schedule. An interesting follow-up experiment could be to examine optimal schedules while varying the relevant features of to-be-learned categories (e.g., similarity of the painting styles themselves).

The present studies—which examined specific category similarity relations using two very different types of categories—provide unique contributions to the current literature on inductive category learning in two ways. First, Carvalho and Goldstone (2014) manipulated within- and between-category similarity together, and Zulkiply and Burt (2012) focused only on between-category discriminations. Even previous studies that used realistic or naturalistic materials, such as artists' painting styles, consisted only of landscape paintings which were constructed with both high within- and between-category similarity together (Carvalho & Goldstone, 2014, Experiment 1; Kang & Pashler, 2012; Kornell & Bjork, 2008). In order to directly test the hypotheses that a blocked study schedule fosters the noticing of within-category commonalities and an interleaved study schedule fosters the noticing of between-category discriminations, we examined category structures with high within- and low between-category similarity and low within- and high-between category similarity. By making the commonalities or the differences trivially easy to spot, these hypotheses made very clear predictions for when blocking was beneficial and when interleaving was beneficial.

Second, prior studies examining category similarity structure (e.g., Carvalho & Goldstone, 2014; Zulkiply & Burt, 2012) utilized artificial categories. While these artificial categories had the advantage of tightly controlling the defining characteristics and features of each category, the present studies extend the research to two types of categories (perceptual-based and conceptual-based) that are more realistic and educationally relevant. Most statistical tests are defined based on several characteristics (e.g., the type of dependent variable or number of groups as independent variables), some of which may overlap with characteristics of other statistical tests (e.g., sample for both independent *t*-test and analysis of variance is unrelated or independent), but the specific combination of defining features can be identified and verbalized for a given test (e.g., number of possible groups is two for independent *t*-test and two or more for analysis of variance). The statistical test categories also differ from the artificial categories used in prior studies in that they are not as perceptually rich in nature, but rather, the search for key features is more conceptual. The artists' paintings that we used are perceptual in nature and different from the artificial stimuli such that they, like many categories in the real world, are not as well defined, and the rules in such cases may not be completely diagnostic of a category membership.

Given the fundamental importance of understanding category learning for enhancing learning and instruction, it is not surprising that the initial, counterintuitive finding by Kornell and Bjork (2008)—that interleaving may be more effective for induction than blocking—has received a lot of interest from both researchers and instructors. It is therefore, also important to understand how broadly this effect can generalize and where its boundary conditions lay. The findings from the present studies not only extend the theoretical framework of optimal sequencing for inductive category learning but, practically, they also demonstrate the generalizability of the interleaving research to educationally relevant and non-artificial categories.

References

Ashby, F.G. & Maddox, W.T. (2011). Human category learning 2.0. *Annals of the NY Academy of Sciences.*

Birnbaum, M. S., Kornell, N., Bjork, E. L., & Bjork, R. A. (2013). Why interleaving enhances inductive learning: The roles of discrimination and retrieval. *Memory & Cognition*, *41*, 392-402.

Carpenter, S. K., & Mueller, F. E. (2013). The effects of interleaving versus blocking on foreign language pronunciation learning. *Memory & Cognition*, *41*, 671-682.

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

Gagné, R. M. (1950). The effect of sequence of presentation of similar items on the learning of paired-associates. *Journal of Experimental Psychology, 40*, 61-73.

Gentner, D., Loewenstein, J., Thompson, L., & Forbus, K. D. (2009). Reviving inert knowledge: Analogical abstraction supports relational retrieval of past events. *Cognitive Science*, *33*, 1343-1382.

Gentner, D., & Namy, L. L. (1999). Comparison in the development of categories. *Cognitive Development*, *14*, 487-513.

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

Kornell, N. & Bjork, R. A. (2008). Learning Concepts and Categories: Is Spacing the "Enemy of Induction"? . *Psychological Science, 19*, 585-592.

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

Kurtz, K. H., & Hovland, C. I. (1956). Concept learning with differing sequences of instances. *Journal of Experimental Psychology*, *51*, 239.

Oakes, L. M., & Ribar, R. J. (2005). A comparison of infants' categorization in paired and successive presentation familiarization tasks. *Infancy*, *7*, 85-98.

Rohrer, D., & Taylor, K. (2007). The shuffling of mathematics practice problems boosts learning. *Instructional Science, 35*, 481–498.

Taylor, K., & Rohrer, D. (2010). The effects of interleaving practice. *Applied Cognitive Psychology, 24*, 837-848.

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

Zulkiply, N., & Burt, J. S. (2013). The exemplar interleaving effect in inductive learning: Moderation by the difficulty of category discriminations. *Memory & Cognition*, *41*, 16-27.

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

Chapter 5: General Discussion

The main goals of the current dissertation were to explore practice schedules that optimize the inductive learning of complex categories, to investigate the cognitive mechanisms that make a given schedule effective for category induction, to explore potential factors that moderate the schedule efficacy, and to examine if the learning advantage conferred by these schedules generalizes across different types of categories and across learners of varying cognitive abilities.

Summary of Results

Generality of the interleaving effect

Findings from this dissertation suggest that the interleaving effect generalizes to the inductive learning of statistical concepts. Participants were better able to identify the statistical concepts of previously unseen test problems on a classification test when study problems were interleaved with problems of other concepts rather than when study problems were blocked by concept (Chapters 1, 2, and 3).

Evidence supporting the discriminative-contrast hypothesis

One possible explanation for the interleaving effect is that juxtaposing exemplars from different categories makes relevant features that discriminate between those categories salient. We examined this discriminative-contrast hypothesis in two ways: by disrupting the contrast processes, and by making the contrast processes more effective.

If the advantage of interleaving is indeed due to discriminative contrast, then disrupting these contrast processes should impair category learning. We examined this prediction of the discriminative contrast hypothesis by inserting 30-sec unrelated fillers (meant to offset the contrast processes) between study problems in the interleaved condition. In support of the hypothesis, participants' classification performance did decrease when the fillers were added, eliminating the interleaving effect (Chapter 3, Experiment 2).

In further support of the discriminative-contrast hypothesis, we predicted that interleaving should produce even greater learning gains under conditions in which learners view exemplars from different categories at once instead of sequentially. This

benefit of simultaneous over sequential study is presumably because the between-category comparisons available to the learner by way of temporal juxtaposition are not as explicit in a sequential study as they may be in a simultaneous schedule, which offers a more explicit learning context to elicit the critical differences. In Chapter 3 (Experiment 3), we found that participants were better able to identify the concepts of previously unseen test problems (although not significantly so) when interleaved study problems were presented three at a time rather than one at a time. Although the benefit was not significant, the direction of performance does provide suggestive evidence for the discriminative-contrast account.

Evidence supporting the study-phase retrieval hypothesis

Another possible explanation for the interleaving effect is the study-phase retrieval hypothesis. Interleaved schedules inherently space out successive exemplars from any given category. We know from the memory literature that spacing is a powerful strategy that enhances long-term learning (Cepeda et al., 2006), and under this hypothesis, the benefits of interleaving for category learning are essentially the same as the benefits of spacing. Largely accepted as the underlying mechanism for the spacing benefit in memory, the study-phase retrieval hypothesis proposes that when spacing is introduced between repetitions (or, in this case, between exemplars from the same category), forgetting of the initial presentation occurs, which leads to subsequent retrieval and reminding of that initial presentation when the next repetition or category exemplar appears. It is the retrieval and/or reminding that effectively solidifies the memory trace and slows down the forgetting rate of the retrieved features. If the advantage of interleaving for category learning is this process of forgetting and retrieval of critical features, then adding temporal spacing to a blocked schedule should also increase classification performance. Indeed, we found that participants were better able to identify the concepts of previously unseen test problems when study problems in the blocked schedule were spaced apart for 30-sec (unrelated filler), replicating the classic spacing effect from memory literature, and from Birnbaum et al. (2013) (Chapter 3, Experiment 2).

Participants' WMC data from this dissertation also corroborate that spacing contributes to the efficacy of study schedules. The WM literature shows that low-WMC individuals encode contextual cues at a global rather than a specific level, which results in increased intrusions and errors at retrieval (Kane & Engle, 2000; Rosen & Engle, 1998). In the present studies, low WMC participants benefited more from a temporally spaced, blocked schedule (compared to an unspaced, blocked schedule). A temporally spaced, blocked schedule may have particularly helped lower WMC participants build several retrieval routes to access the critical features of a given category, and consequently foster an accumulation of more specific and relevant cues that they can later use to ensure successful retrieval (Brewer & Unsworth, 2012; Tse & Pu, 2012).

According to the study-phase retrieval hypothesis, the role of retrieval difficulty can become a hindrance to learning if too much temporal spacing or too many disruptions between exemplars make it impossible to retrieve previously studied features. In support of this notion, low WMC participants, who were limited in their ability to successfully retrieve previous, did not benefit from increasing temporal spacing between study problems in the interleaved schedules presumably because the disruptive fillers between problems—in addition to the increased memory load from learning multiple artists at the same time—made retrieving the features too difficult or unsuccessful. We know from prior WM research that at retrieval, individuals use internally generated contextual cues (e.g., recall features only from the previous two exemplars) to delimit their search set to include relevant features (e.g., features specific to the current category) and to exclude irrelevant features (e.g., features from different categories) (Rosen & Engle, 1998). Lower WMC individuals demonstrate noisier internally generated context cues, which leads to poorer recall of the target features (Unsworth & Engle, 2007). Interpreting the results of Chapter 3, Experiment 2 in light of this literature, inserting disruptions in the form of temporal spacing in an interleaved schedule, which already produces high contextual interference, may have enabled retrieval attempts based on both irrelevant and relevant category cues.

Category similarity structure moderating the interleaving effect

We demonstrate that each of these two schedules, blocked and interleaved, foster unique category comparisons: an interleaved schedule leads to better induction when there is high between-category similarity and low within-category similarity (i.e., where noticing differences between categories is critical) as predicted by the discriminative-contrast hypothesis; a blocked schedule leads to better induction when there is low between-category similarity and high within-category similarity (i.e., where noticing the similarities within-categories is critical). With the conceptual-based categories (i.e., the statistical concepts), we found that participants were better able to identify the concepts of previously unseen test problems in the interleaved condition when content of the to-be-learned category sets overlapped (i.e., emphasized between-category comparisons), and in the blocked condition when content of each category set was distinct (i.e., fostered within-category comparisons) (Chapter 4, Experiment 1). Also in line with the notion that the benefit of category learning depends on the salient characteristics that a schedule emphasizes, an interleaved schedule can be made more or less effective by manipulating the specific juxtapositions of the content of the categories—when comparing interleaved schedules, massing the content of study problems allowed participants to more easily disregard the irrelevant features and focus instead on the critical differences between categories compared to shuffling the content of the study problems (Chapter 4, Experiment 3A).

In Chapter 4 (Experiment 2), with the perceptual-based categories (i.e., the artistic painting styles), an interleaved schedule led to better induction when content of the to-be-learned category sets overlapped (i.e., fostered between-category comparisons), as proposed by the discriminative-contrast hypothesis, and as shown with our conceptual categories above. Contrary to our predictions, however, a blocked schedule did not lead to better induction when content of each category set was distinct (i.e., fostered within-category comparisons), although the interleaving advantage was eliminated. In Experiment 4 (Chapter 4), with the content of each artist's paintings distinct from the content of the other artists, we created a situation in which the blocked schedule should

have produced a greater advantage over the interleaved schedule. We did this by testing learners' on their ability to distinguish between paintings by studied artists and paintings by unstudied artists with similar content (i.e., lures). We predicted that interleaved schedules would draw learners' attention to key differences (i.e., the content) between the artists, whereas blocked schedules would draw their attention to the similarities between an artist's paintings (i.e., their painting style beyond just the superficial content details). Therefore, the blocked schedule should lead to better discrimination between the studied artists' paintings and the content-lures. No such difference in classification test performance, however, was observed. Similarly, in Experiment 3B (Chapter 4), where the overlapped artists' paintings were presented interleaved, classification performance was similar regardless of whether the content of exemplars was massed or shuffled, despite our predictions that massing the content should reduce superficial noise and highlight critical differences between artists (as was found in Experiment 3A with the conceptual categories).

Nature of categories moderating the interleaving effect

The discrepancy in the interaction between study schedule and category similarity structure may partly appear to be because of the nature of the two sets of categories: conceptual, rule-based (i.e., statistical concepts) and perceptual-based (artists' painting styles). When there was low between-category similarity and high within-category similarity, we observed a blocking benefit, as predicted, with the conceptual categories but not with the perceptual categories. Even when we set up conditions in which a blocked schedule should have increased performance more than an interleaved schedule, we found no difference in performance with the perceptual categories. In the current dissertation, we cannot isolate what precisely it is about the nature of the categories that leads to such diverging pattern of results, but one potential candidate is the dual-process framework of category induction (Ashby & Maddox, 2011), which differentiates between rule-based learning and information-integration learning.

Rule-based learning is thought to be an explicit process in which learning is best supported by a strategy of hypothesis-testing and rule discovery. In the prior studies

using artificial stimuli (e.g., Carvalho & Goldstone, 2014; Zulkiply & Burt, 2012) and in our study (Chapter 4, Experiment 1) using conceptual-based stimuli (i.e., statistical tests), the categories are well defined by specific rules that learners must discover. With these well-defined, rule-based categories, it appears that category similarity structure mediates schedule effects, such that interleaving outperforms blocking when between-category similarity is high, and blocking outperforms interleaving when between-category similarity is low. On the other hand, information-integration learning is thought to be an implicit process and facilitates learning in cases where the rules are not verbalizable—our artists' painting styles categories may fall more in line with this type of learning (although, these materials are not purely information-integration categories; we acknowledge that there are components of these artists paintings that may be somewhat verbalizable). Category similarity structure may not mediate schedule effects to the same extent for information-integration category learning, in part, potentially because blocking encourages explicit hypothesis-testing (which is not optimal for implicit, information-integration learning) while interleaving makes such explicit hypothesis-testing difficult. In short, these data suggest that perhaps the nature of the categories themselves may drive induction differently, and moderate the interleaving benefit.

Temporal spacing moderating the interleaving effect

In addition to the category similarity structure and the nature of categories, temporal spacing moderates the benefit of interleaving, and can even enhance learning in blocked schedules. In Chapter 3, Experiment 2, we demonstrated that increasing temporal spacing between exemplars in an interleaved schedule eliminates the interleaving effect (presumably because the fillers disrupt contrast processes, in support of the discriminative-contrast hypothesis) while introducing these fillers in a blocked schedule produces a spacing benefit (in support of the study-phase retrieval hypothesis). We also found that eliminating temporal spacing by presenting study problems simultaneously rather than sequentially not only improved classification performance in the interleaved schedules, as would be predicted by the discriminative-contrast hypothesis, but also improved classification performance in the blocked schedules (Chapter 3, Experiment 2).

At first glance, the two results appear to contradict one another—learning in blocked schedules is improved by both increasing spacing (Chapter 3, Experiment 2) and eliminating spacing (Chapter 3, Experiment 3). One way of potentially reconciling these findings is that increasing spacing enhances learning by encouraging more difficult (but still successful) retrieval of concept features, whereas eliminating spacing enhances learning by encouraging greater comparison of and attention to common, relevant features. Thus, the two processes that enhance blocked learning here may be different. Although there is converging evidence for the robust effects of spaced learning (Cepeda et al., 2006, 2008, 2009), there is also evidence to suggest that simultaneously presenting problems may foster more direct comparisons, which in turn helps learners overcome contextual limitations and allow them to recognize the common deep features (e.g., Catrambone & Holyoak, 1989; Markman & Gentner, 2005). Moreover, presenting learners with problems that differ in their surface features (e.g., cover stories, names, objects of the study problems) but share similar structural features (e.g., principles, equations, and procedures) can further enable this comparison process as their attention is more quickly directed to what features are and what features are not relevant for categorization (e.g., Quilici & Mayer, 1996, 2002).

Working memory capacity moderating the interleaving effect

Finally, another potential moderator of the interleaving benefit is the WMC of individual learners, which could influence selective attention and processing of relevant incoming information and its integration with existing information in LTM (e.g., Unsworth & Engle, 2007; Unsworth & Spillers, 2010). We initially predicted that low-WMC individuals would benefit less from interleaving because interleaved study places greater WM demands on the learner than does blocked study. We found, however, that interleaving exemplars from multiple categories was significantly more effective than blocking by category for the inductive learning of both low- and high-WMC participants, and this effect held with both perceptual-based categories (artists' painting styles; Chapter 2, Experiment 1) and conceptual-based categories (statistical tests; Chapter 2, Experiment 2; Chapter 3, Experiments 1–3). In fact, interleaving appeared to produce greater learning

gains for participants with low WMC than for participants with high WMC (at least for the of statistical concepts). Why did these participants benefit from interleaved schedules when there were compelling theoretical reasons to expect otherwise? Perhaps the enhanced discriminative contrast associated with interleaving increases the salience of relevant features across categories (a comparison opportunity not afforded in blocked practice), alleviating limitations associated with having low WMC.

The classification performance differences between the high- and low-WMC participants were greatly reduced not only with the interleaved schedules, but also with the blocked schedule for which we increased temporal spacing (Chapter 3, Experiment 2). According to the study-phase retrieval hypothesis, a typical blocked condition requires very little retrieval and minimal forgetting of previous problem features. Conversely, a blocked schedule with spacing may be more effective because the temporal spacing causes forgetting of previous problem features, which then requires that participants retrieve those features from LTM. Given that their cue-dependent controlled search from LTM is encoded at a global rather than a specific level, which results in increased intrusions and errors (Kane & Engle, 2000; Rosen & Engle, 1998), this schedule may help lower WMC participants to build several retrieval routes, and consequently foster an accumulation of more specific cues (Brewer & Unsworth, 2012; Tse & Pu, 2012).

Our findings also suggest that disrupting the contrast processes in the interleaved schedule (through insertion of unrelated fillers between study problems) impaired classification performance, particularly for low-WMC participants (Chapter 3, Experiment 2). Making comparisons across categories would be more difficult in interleaved schedules with temporal spacing, particularly for this group who already have difficulty controlling their attention by maintaining focus on relevant information (e.g., discriminative features between categories) and inhibiting interference from distraction (e.g., distraction caused by the 30-sec fillers) (Kane & Engle, 2000; Rosen & Engle, 1998). We can also interpret the WMC results through a "desirable difficulties" (Bjork, 1994) perspective, which suggests that learning conditions that are more challenging during study often promote more elaborative encoding and optimize long-term retention.

An individual's WMC, however, likely determines the optimal level of difficulty for learning. Those with lower WMC are less able than those with higher WMC to actively maintain and process task-relevant information and retrieve related information from LTM in the face of distraction (Engle & Kane, 2004; Unsworth & Spillers, 2010). Thus, the threshold for difficulty being desirable rather than undesirable (i.e., inducing cognitive overload) may be lower for learners with lower WMC. In other words, lower WMC learners may be more susceptible to practice schedules that tax their WM.

Theoretical Contributions and Practical Implications

Findings from this dissertation provide unique contributions to the cognitive and educational research.

Providing further evidence for the role of discriminative contrast

As reported in Chapter 3, we tried to dissociate the two hypotheses: discriminative-contrast and study-phase retrieval, by manipulating temporal spacing. While other studies have also attempted to test these hypotheses, the results, so far, have been mixed. Taylor and Rohrer (2010, mathematics formulae) and Kang and Pashler (2012, artists' painting styles) controlled for spacing by comparing an interleaved schedule with a blocked (temporally spaced) schedule. They still found an interleaving benefit, leading them to conclude that the benefits of interleaving arise from directly juxtaposing exemplars of different categories. Neither of these studies, however, compared classification performance between interleaved schedules with and without temporal spacing to more directly test the discriminative-contrast hypothesis, whereas Birnbaum et al. (2013, butterflies' species) and Zulkiply and Burt (2012, artists' painting styles) did. Birnbaum et al. found that performance in the interleaved condition decreased when fillers were added to disrupt contrast processes, whereas Zulkipy and Burt did not find this result (although adding spacing to an interleaved schedule did numerically decrease classification performance). Findings from the current dissertation are consistent with those from Birnbaum et al.—the interleaving benefit was eliminated when the fillers were added to the schedule, presumably because the fillers interrupted the contrast processes necessary to obtain an interleaving benefit (Chapter 3, Experiment 2,

statistical tests). This finding provides clear evidence for the role of category comparisons in the benefit of interleaving given that comparisons would be more difficult in interleaved schedules with temporal spacing.

Demonstrating blocking benefits

The findings in this dissertation demonstrate that blocked schedules may be equally effective when learning textual, rule-based conceptual categories if they promote effortful, but successful retrieval (via temporal spacing), and foster within-category comparisons (via simultaneous presentations and via low between- and high within-category similarity structure). Whereas, Kang and Pashler (2012) and Zulkiply and Burt (2012) did not find a spacing effect when they compared blocked conditions with and without temporal spacing, Birnbaum et al. (2013) and the present study from Chapter 3 (Experiment 2) did observe a blocking benefit, a finding that is consistent with the study-phase retrieval hypothesis. Perhaps the different pattern of results may be because, unlike learning artists' painting styles (Kang & Pashler, 2012; Zulkiply & Burt, 2012), characteristics that define different butterflies' are verbalizable (e.g., specific patterns, shapes and colours of the wings), similar to the rule-based statistical concepts; blocked schedules encourage explicit hypothesis testing, which is more favourable to rule-based or feature-based categories.

With regard to enhancing similarities within a category, Kang and Pashler (2012, artists' painting styles) and Wahlheim et al. (2011, birds' families) compared blocked conditions in which exemplars were presented one at a time or together (either all at once or in pairs). They found no additional benefit of presenting exemplars simultaneously rather than sequentially, whereas findings from Chapter 3 (Experiment 3) did show the additional benefit. One possible reason for the different pattern of results across the studies may again be due to the nature of categories. Compared to perceptual categories, statistical concepts are not only rule-based, and thus, enable explicit hypothesis testing, but they also have contextual information in the problems that can, during simultaneous presentations, further enable the comparison process by emphasizing relevant features (e.g., Quilici & Mayer, 1996). In fact, studies on analogical transfer have shown that

learners do not spontaneously engage in comparisons unless multiple exemplars of a topic are studied simultaneously rather than studied separately (e.g., Gentner et al., 2003).

The majority of studies examining the relative effects of blocking versus interleaving have used low-discriminability categories—that is, categories that are all very similar to one another. The findings from Chapter 4 (Experiment 1) demonstrate that changing the similarity structure of categories such that they are highly discriminable—that is, where all the exemplars from the same category are highly dissimilar and the few features they share are difficult to identify—can reverse the benefit of interleaving, and in fact, promote learning in a blocked schedule.

Manipulating category similarity structure

When examining the category similarity structure, Carvalho and Goldstone (2014) manipulated within- and between-category similarity together, and Zulkiply and Burt (2012) focused only on between-category discriminations. Even previous studies that used realistic or naturalistic materials, such as artists' painting styles, consisted only of landscape paintings which were constructed with both high within- and between-category similarity together (Carvalho & Goldstone, 2014, Experiment 1; Kang & Pashler, 2012; Kornell & Bjork, 2008). In order to directly test the hypotheses that a blocked study schedule fosters the noticing of within-category commonalities and an interleaved study schedule fosters the noticing of between-category discriminations, we examined category structures with high within- and low between-category similarity and low within- and high-between category similarity. By making the commonalities or the differences trivially easy to spot, these hypotheses made very clear predictions for when blocking was beneficial and when interleaving was beneficial. Consistent with these predictions, we demonstrate that the efficacy of a given schedule does indeed seem to depend on the similarity of the category structures.

Examining individual differences in learners' WMC

Relatively comprehensive sets of instructional methods, including interleaved practice have been established to optimize learning (Dunlosky et al., 2013). Although these methods are, without a doubt, helpful in terms of planning instruction and

developing learning materials through an understanding of the learners' information processing capacity, the potential interactions between the methods and learners' WMC are not well understood. Does a given method provide general benefits across all ability ranges, such that all students benefit equally? Does it preferentially help good students to better utilize their inherent abilities, but does little to help other students? Does it minimize differences in performance across the ability range, such that it can be uniformly applied in the classroom? As reported in Chapters 2 and 3 of this dissertation, we found that all learners likely benefited from the optimal practice schedules (i.e., interleaving without temporal spacing and blocking with temporal spacing), but low ability learners benefited most, presumably because those schedules encouraged engagement of cognitive abilities not normally employed. These results on individual differences in WMC provide a theoretical basis for tailoring learning and instruction, particularly practice schedules for the learning of categories and concepts, and demonstrate the generalizability of interleaving benefits across different learner abilities.

Using ecologically valid materials

Most of the prior studies, with the exception of Rohrer and colleagues (math stimuli) and Bjork and colleagues (perceptual, realistic stimuli), utilized artificial categories (e.g., Carvalho & Goldstone, 2014; Zulkiply & Burt, 2012), which may have the advantage of tightly controlling the defining characteristics of each category, but their findings may not generalize to other types of more realistic and educationally relevant categories. The artists' paintings that we used are perceptual in nature and different from the artificial stimuli such that they, like many categories in the real world, are not as well defined, and the rules in such cases may not be completely diagnostic of a category membership. The statistical tests that we used differ both from the artificial and perceptual stimuli in that they are not as perceptually rich in nature, and the search for key features is more conceptual—most statistical tests are defined based on several features (e.g., the type of dependent variable or number of independent variables), some of which may overlap with features of other statistical tests (e.g., sample for both independent *t*-test and analysis of variance is unrelated), but the specific combination of

defining features can be identified and verbalized for a given concept. The use of such ecologically valid stimuli gives us some confidence that the interleaving effect has the potential to generalize to other similar kinds of challenging, educationally relevant concepts.

Applying results to educational settings

Finding from this dissertation show that properly constructed learning opportunities (e.g., scheduling practice or spacing study sessions) can promote knowledge acquisition. There are several straightforward practical ways in which these opportunities can be introduced in classrooms and in instructional books. For instance, assignments and end-of-chapter problems might include problems not just from the current topic, but also intermix problems from previous units or chapters. This interleaved arrangement offers students the chance to practice mapping exemplar problems to their corresponding concepts, and thereby gain the knowledge they need to make discriminations across concepts. Instructors in classrooms can also lead explicit discussions on the commonalities and differences across concepts. This direct instruction may offer benefits not only in terms of increasing students' understanding, but also in terms of encouraging students to search out comparison opportunities on their own.

Future Studies

With regard to the theoretical framework of the interleaving effect, the evidence suggests that both discriminative-contrast and retrieval processes play a role. The supporting evidence for the discriminative-contrast hypothesis comes from manipulations that disrupt the contrast processes critical to interleaving (chapter 3), or that shift the focus away from between-category comparisons (chapter 4). The supporting evidence for the study-phase retrieval hypothesis comes particularly when learners' WMC and its relation with retrieval processes are examined (chapter 3). Thus, the issue is not really about which of the two accounts better explains the interleaving effect, as both play a role, but rather about examining the conditions under which one account may play a greater role than the other. For example, if there are more categories that need to be learned, the processes related to retrieval difficulty may become more important, and if

the categories get harder to discriminate, the processes related to discrimination become more important.

Moreover, much research remains to be done before we fully understand the boundary conditions of the interleaving effect, particularly with respect to how it may be affected by the complexity of the task. Some research suggests that blocking may be more effective than interleaving when the learning task is particularly difficult. For example, de Croock and van Merrieenboer (2007) presented participants with different types of problems that could occur in a simulated complex distiller system (e.g., pipe leakage, sensory malfunction, etc.), and asked them to troubleshoot these problems. Participants performed better when problems were presented to them in a way that was blocked by type of malfunction, rather than interleaved such that the type of malfunction was different on each problem. For concepts that are intrinsically more difficult to learn, it may also be worth exploring whether a mixture of blocking and interleaving is optimal. Rather than using a schedule that is exclusively blocked or interleaved, it may be more advantageous to start with a blocked schedule and then transition to interleaving when learning difficult concepts. This possibility has been discussed in other papers (see, e.g., Dunlosky et al., 2013; Rohrer, 2012) but has yet to be fully explored (but see Yan, 2014 who reported superior classification performance when learning artists' painting styles with a schedule that was blocked-to-interleaved rather than interleaved-to-blocked).

There are several other research questions that remain underexplored. For instance, we found a sub-additive interaction between interleaving and spacing (Chapter 3, Experiment 2). If we are to apply cognitive research to the classroom, it is not only important to understand a given instructional method on its own, but it is also important to understand how the method interacts in conjunction with other instructional methods. For instance, participants who explicitly generate comparison-based explanations (generation being a robust learning method; Chi, 2000) during interleaved versus blocked schedules may recognize faster, and encode better the critical features of to-be-learned categories. Though such an account is admittedly a speculative one at present, examining it can provide insight into the kind of information that participants attend to, and the form

of explicit associations that they make across the different practice schedules.

Another important area to further investigate is based on the underlying theme that is consistent across the several experiments in this dissertation—failure to replicate interactions between WMC and practice schedules (Chapter 2) and between practice schedules and category similarity structure (Chapter 4) across two very different categories: artists' paintings vs. statistical tests. It seems that the nature of the categories themselves may drive induction differently. This highlights the need to further explore how the different types of categories (e.g., categories that are perceptual-based versus conceptual-based) map on to the dual-process framework of category induction (i.e., category learning that is rule-based versus information-integration-based).

Concluding Remarks

The ability to discriminate between categories so that one can recognize and classify new exemplars of those categories is a fundamental process that underlies much of education and learning. A better understanding how to enhance category learning will allow us to tailor instruction depending on the to-be-learned categories and depending on the learners' cognitive abilities. Prior studies have demonstrated that counter to most learners' intuitions (and certainly in contrast to how formal education is typically arranged), intermixing the learning of multiple, related categories (i.e., interleaved schedule) is more effective for category learning than is focusing the learning on one category at a time. The series of experiments reported here demonstrate that the interleaving benefit is a general effect: the findings show that the interleaving benefits extend across different content domains (art and statistics) and across different learner abilities (low- and high- WMC learners). As for the theoretical framework of the interleaving effect, both discriminative-contrast and retrieval processes play a role, one more than other depending on, for example, category discriminability and retrieval difficulty. And finally, there are several moderators that contribute to the relative efficacy of blocked and interleaved schedules, including category similarity structure, different types of categories, and temporal spacing within and between categories. Overall, findings from the current dissertation not only extend the theoretical framework of

optimal study schedules for inductive category learning to consider multiple factors, but it also has practical importance for generalizing the research to educationally-relevant and non-artificial categories.

References

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

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. doi: 10.1016/j.jecp.2009.11.003

Ashby, F.G. & Maddox, W.T. (2011). Human category learning 2.0. *Annals of the NY Academy of Sciences.*

Baddeley, A.D., & Hitch, G.J. (1974). Working memory. In G. H. Bower (Ed.), *The psychology of learning and motivation, Vol. 8* (pp. 47-89). New York: Academic Press.

Benjamin, A.S. & Ross, B.H. (2011). The causes and consequences of reminding. In A. S. Benjamin (Ed.), *Successful remembering and successful forgetting: A Festschrift in honor of Robert A. Bjork* (pp. 71–88). New York, NY: Psychology Press.

Birnbaum, M. S., Kornell, N., Bjork, E. L., & Bjork, R. A. (2013). Why interleaving enhances inductive learning: The roles of discrimination and retrieval. *Memory & Cognition*, *41*, 392-402.

Bjork, R.A. (1975). Retrieval as a memory modifier. In R. Solso (Ed.), *Information processing and cognition: The Loyola Symposium* (pp. 123–144). Hillsdale, NJ: Lawrence Erlbaum Associates.

Bortoli, L., Robazza, C., Durigon, V., & Carra, C. (1992). Effects of contextual interference on learning technical sports skills. *Perceptual and Motor Skills*, *75*(2), 555–562. doi: 10.2466/pms.1992.75.2.555

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

Capaldi, E.J., & Neath, I. (1995). Remembering and forgetting as context discrimination. *Learning & Memory*, *2*(3-4), 107–132. doi: 10.1101/lm.2.3-4.107

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

Carvalho, P.F. & Goldstone, R.L. (2015). What you learn is more than what you see: What can sequence effects tell us about inductive category learning? *Frontiers in Psychology, 6*(505). doi: 10.3389/fpsyg.2015.00505

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

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

Catrambone, R., & Holyoak, K.J. (1989). Overcoming contextual limitations on problem-solving transfer. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, *15*(6), 1147–1156. doi: 10.1037/0278-7393.15.6.1147

Cepeda, N.J., Coburn, N., Rohrer, D., Wixted, J.T., Mozer, M.C., & Pashler, H. (2009). Optimizing distributed practice: theoretical analysis and practical implications. *Experimental Psychology, 56*(4), 236–246. doi: 0.1027/1618-3169.56.4.236

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

Cepeda, N.J., Vul, E., Rohrer, D., Wixted, J.T., & Pashler, H. (2008). Spacing effects in learning: a temporal ridgeline of optimal retention. *Psychological Science, 19*(11)*,* 1095–1102. doi: 10.1111/j.1467-9280.2008.02209.x

Conway, A.R., 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. doi: 10.3758/BF03196772

Dempster, F.N. (1988). The spacing effect: A case study in the failure to apply the results of psychological research. *American Psychologist*, *43*(8), 627-634. doi:

10.1037/0003-066X.43.8.627

Engle, R.W., & Kane, M.J. (2004). Executive attention, working memory capacity, and a two-factor theory of cognitive control. In B. Ross (Ed.), *The psychology of learning and motivation, Vol. 44* (pp. 145-199). New York: Elsevier.

Gagné, R. M. (1950). The effect of sequence of presentation of similar items on the learning of paired-associates. *Journal of Experimental Psychology, 40*, 61-73.

Gentner, D. (1983). Structure-mapping: A theoretical framework for analogy. *Cognitive Science*, *7*(2)*,* 155–170. doi: 10.1207/s15516709cog0702_3

Gentner, D., & Namy, L. L. (1999). Comparison in the development of categories. *Cognitive Development*, *14*, 487-513.

Gentner, D., Loewenstein, J., & Thompson L. (2003). Learning and transfer: A general role for analogical encoding. *Journal of Educational Psychology, 95*(2), 393– 408. doi: 10.1037/0022-0663.95.2.393

Gentner, D., Loewenstein, J., Thompson, L., & Forbus, K.D. (2009). Reviving inert knowledge: Analogical abstraction supports relational retrieval of past events. *Cognitive Science*, *33*(8), 1343-1382. doi: 10.1111/j.1551-6709.2009.01070.x

Gick, M.L., & Holyoak, K.J. (1983). Schema induction and analogical transfer. *Cognitive Psychology, 15*(1)*,* 1-38. doi: 10.1016/0010-0285(83)90002-6

Gick, M.L., & Holyoak, K.J. (1987). The cognitive basis of knowledge transfer. In S. M. Cormier & J. D. Hagman (Eds.), *Transfer of learning: Contemporary research and applications* (pp. 9-46). Orlando, FL: Academic Press.

Goldstone, R.L. (1996). Isolated and interrelated concepts. *Memory & Cognition. 24*(5)*,* 608–628. doi: 10.3758/BF03201087

Goldstone, R.L., & Steyvers, M. (2001). The sensitization and differentiation of dimensions during category learning. *Journal of Experimental Psychology: General*, *130*(1), 116–139. doi: 10.1037/0096-3445.130.1.116

Goode, S., & Magill, R.A. (1986). Contextual interference effects in learning three badminton serves. *Research Quarterly for Exercise and Sport*, *57*(4), 308–314. doi: 10.1080/02701367.1986.10608091

Gravetter, F.J., & Wallnau, L.B. (2008) *Statistics for the behavioral sciences* (8th ed.). Belmont, CA: Wadsworth Cengage Learning.

Hall, K.G., Domingues, D.A., & Cavazos, R. (1994). Contextual interference effects with skilled baseball players. *Perceptual and Motor Skills*, *78*(3), 835–841. doi: 10.2466/pms.1994.78.3.835

Hintzman, D.L. (2004). Judgment of frequency versus recognition confidence: Repetition and recursive reminding. *Memory & Cognition*, *32*(2), 336–350. doi: 10.3758/BF03196863

Kane, M.J., & Engle, R.W. (2000). Working-memory capacity, proactive interference, and divided attention: Limits on long-term memory retrieval. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, *26*(2), 336-358. doi: 10.1037/0278-7393.26.2.336

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

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

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

Kornell, N., Castel, A.D., Eich, T.S., & Bjork, R.A. (2010). Spacing as the friend of both memory and induction in young and older adults. *Psychology and Aging*, *25*(2), 498–503. doi: 10.1037/a0017807

Kurtz, K.H., and Hovland, C.I. (1956). Concept learning with differing sequences of instances. *Journal of Experimental Psychology, 51*(4), 239–243. doi: 10.1037/h0040295

Kurtz, K.J., Miao, C.H., & Gentner, D. (2001). Learning by analogical bootstrapping. *The Journal of the Learning Sciences*, *10*(4), 417-446. doi: 10.1207/S15327809JLS1004new_2

Le Blanc, K., & Simon, D. (2008, November). Mixed practice enhances retention and JOL accuracy for mathematical skills. In *49th Annual Meeting of the Psychonomic Society, Chicago, IL*.

Le Blanc, K., & Simon, D. (2008, November). Mixed practice enhances retention and JOL accuracy for mathematical skills. In *49th Annual Meeting of the Psychonomic Society, Chicago, IL*.

Lewandowsky, S. (2011). Working memory capacity and categorization: Individual differences and modeling. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *37*(3), 720–738. doi: 10.1037/a0022639

Markman, A.B., & Gentner, D. (2005). Nonintentional similarity processing. In R. Hassin, J.A. Bargh and J.S. Uleman (Eds.), *The new unconscious* (pp. 107-137). New York: Oxford University Press.

Mayer, R.E., Sims, V., & Tajika, H. (1995). A comparison of how textbooks teach mathematical problem solving in Japan and the United States. *American Educational Research Journal, 32*(2)*,* 443–460. doi: 10.3102/00028312032002443

Mayfield, K.H., & Chase, P.N. (2002). The effects of cumulative practice on mathematics problem solving. *Journal of Applied Behavior Analysis*, *35*(2), 105-123. doi: 10.1901/jaba.2002.35-105

Mundy, M.E., Honey, R.C., and Dwyer, D.M. (2007). Simultaneous presentation of similar stimuli produces perceptual learning in human picture processing. *Journal of Experimental Psychology: Animal Behavior Processes, 33*(2), 124–138. doi: 10.1037/0097- 7403.33.2.124

Oakes, L. M., & Ribar, R. J. (2005). A comparison of infants' categorization in paired and successive presentation familiarization tasks. *Infancy*, *7*, 85-98.

Ollis, S., Button, C., & Fairweather, M. (2005). The influence of professional expertise and task complexity upon the potency of the contextual interference effect. *Acta*

Psychologica, *118*(3), 229–244. doi: 10.1016/j.actpsy.2004.08.003

Quilici, J.L., & Mayer, R.E. (1996). Role of examples in how students learn to categorize statistics word problems. *Journal of Educational Psychology*, *88*(1), 144–161. doi: 10.1037/0022-0663.88.1.144

Quilici, J.L., & Mayer, R.E. (2002). Teaching students to recognize structural similarities between statistics word problems. *Applied Cognitive Psychology*, *16*(3), 325-342. doi: 10.1002/acp.796

Rohrer, D. (2009). The effects of spacing and mixing practice problems. *Journal for Research in Mathematics Education, 40*(1), 4-17. Retrieved from http://www.jstor.org/stable/40539318

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

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

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

Rosen, V.M., & Engle, R.W. (1998). Working memory capacity and suppression. *Journal of Memory & Language*, *39*(3), 418-436. doi: 10.1006/jmla.1998.2590

Sanchez, C.A., & Wiley, J. (2006). An examination of the seductive details effect in terms of working memory capacity. *Memory & Cognition, 34*, 344–355. doi: 10.3758/BF03193412

Sanchez, C.A., & Wiley, J. (2009). To scroll or not to scroll: Scrolling, working memory capacity, and comprehending complex texts. *Human Factors: Journal of the Human Factors and Ergonomics Society*, *51*(5), 730–738. doi: 10.1177/0018720809352788

Shea, J.B., & Morgan, R.L. (1979). Contextual interference effects on the acquisition, retention, and transfer of a motor skill. *Journal of Experimental Psychology: Human Learning and Memory, 5,* 179–187. doi: 10.1037/0278-7393.5.2.179

Shea, J.B., & Zimny, S.T. (1983). Context effects in memory and learning movement information. In R. A. Magill (Ed.), *Memory and control of action* (pp. 345–366). Amsterdam: Elsevier. doi: 10.1016/S0166-4115(08)61998-6

Simon, D.A., & Bjork, R.A. (2001). Metacognition in motor learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *27*(4), 907–912. doi: 10.1037/0278-7393.27.4.907

Star, J.R., & Rittle-Johnson, B. (2009). It pays to compare: An experimental study on computational estimation. *Journal of Experimental Child Psychology*, *102*(4), 408-426. doi: 10.1016/j.jecp.2008.11.004

Sweller, J. (1994). Cognitive load theory, learning difficulty, and instructional design. *Learning and Instruction*, *4*(4), 295–312. doi: 10.1016/0959-4752(94)90003-5

Sweller, J., Van Merrienboer, J.J., & Paas, F.G. (1998). Cognitive architecture and instructional design. *Educational Psychology Review*, *10*(3), 251–296. doi: 10.1023/A:1022193728205

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

Thios, S.J., & D'Agostino, P.R. (1976). Effects of repetition as a function of study-phase retrieval. *Journal of Verbal Learning & Verbal Behavior*, *15*(5), 529-536. doi: 10.1016/0022-5371(76)90047-5

Tse, C.S., & Pu, X. (2012). The effectiveness of test-enhanced learning depends on trait test anxiety and working memory capacity. *Journal of Experimental Psychology: Applied*, *18*(3), 253-264. doi: 10.1037/a0029190

Underwood, B.J. (1957). Interference and forgetting. *Psychological Review, 64*(1), 49–60. doi: 10.1037/h0044616

Unsworth, N., & Engle, R.W. (2007). The nature of individual differences in working memory capacity: active maintenance in primary memory and controlled search from secondary memory. *Psychological Review*, *114*(1), 104-132. doi: 10.1037/0033-295X.114.1.104

Unsworth, N., & Spillers, G.J. (2010). Variation in working memory capacity and

episodic recall: The contributions of strategic encoding and contextual retrieval. *Psychonomic Bulletin & Review*, *17*(2), 200-205. doi: 10.3758/PBR.17.2.200

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. doi: 10.3758/BF03192720

Unsworth, N., Spillers, G.J., & Brewer, G.A. (2012). Working memory capacity and retrieval limitations from long-term memory: An examination of differences in accessibility. *Quarterly Journal of Experimental Psychology*, *65*(12), 2397–2410. doi: 10.1080/17470218.2012.690438

Vlach, H. A., Sandhofer, C. M., & Kornell, N. (2008). The spacing effect in children's memory and category induction. *Cognition*, *109*(1), 163–167. doi: 10.1016/j.cognition.2008.07.013

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

Whitman, J.R., & Garner, W.R. (1963). Concept learning as a function of form of internal structure. *Journal of Verbal Learning & Verbal Behavior*, *2*(2), 195-202. doi: 10.1016/S0022-5371(63)80085-7

Yan, V.X. (2014). *Learning Concepts and Categories from Examples: How Learners' Beliefs Match and Mismatch the Empirical Evidence*. (Unpublished doctoral dissertation). University of California, Los Angeles, U.S.A.

Zulkiply, N., & Burt, J.S. (2012). The exemplar interleaving effect in inductive learning: Moderation by the difficulty of category discriminations. *Memory & Cognition*, *41*(1), 16-27. doi: 10.3758/s13421-012-0238-9

Zulkiply, N., McLean, J., Burt, J.S., & Bath, D. (2012). Spacing and induction: Application to exemplars presented as auditory and visual text. *Learning and Instruction, 22*, 215–221. doi: 10.1016/j.learninstruc.2011.11.00