HOW STIMULUS SIMILARITY IMPACTS SPACING AND INTERLEAVING EFFECTS IN LONG-TERM MEMORY

A DISSERTATION
SUBMITTED TO THE FACULTY OF THE GRADUATE SCHOOL
OF THE UNIVERSITY OF MINNESOTA
BY

Katrina B. Archambault

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

Chad J. Marsolek, Adviser

June 2014

Acknowledgements

This thesis could not have been completed without the support of my advisor, colleagues, committee, family, and friends. First and foremost, I would like to thank my adviser, Chad Marsolek, for providing the intellectual support needed to complete this dissertation and the graduate program in psychology. I am grateful to have worked with someone who is as passionate about understanding the mind as Chad is. I also appreciate Chad's approachability and informality, and have learned that important discussions about research can happen at a concert or over a beer.

I would like to thank Sashank Varma, who has given a large amount of time and effort towards my development as a researcher despite not technically being my adviser, and is largely responsible for my continuing interest in the world of education research. I extend gratitude to the other members of my dissertation committee, Wilma Koutstaal, and Randy Fletcher, for their valuable comments and guidance in the development of this research project.

I would not be graduating if it were not for the emotional support of fellow lab members with whom I have shared the ups and downs of graduate school, Tyler Yost, Vaughn Steele, Alvina Kittur, Susan Park Anderson, Brenton McMenamin, and Mike Blank. I would also like to extend my thanks to the members of Textgroup not already mentioned, especially Brooke Lea and Andrew Elfenbein, who together with the rest of the group have provided a nurturing environment for lively discussion of research that has greatly contributed to my development as a scholar.

I am indebted to the Minnesota Interdisciplinary Training in Education Research (MITER) program for financial support in my first four years of graduate

school, as well as the University of Minnesota Department of Psychology for their support in my most recent years in the program. As part of MITER I was introduced to Nicole Landi, whom I would like to thank for her mentorship in using ERP methodology in education research.

Lastly, I would like to thank my parents. Their confidence in my ability to do whatever I put my mind to has helped lead me to where I am now.

Abstract

In this study, three experiments examined the impact of stimulus similarity on the benefits of spacing and interleaving for long-term memory. Two laboratory-based experiments (Experiments 1 and 2) and one classroom-based experiment (Experiment 3) were conducted. In Experiment 1, an advantage for interleaving relative to massing stimuli during encoding was observed as a greater proportion of correct responses on a categorization test for birds and paintings. This advantage was significantly greater when the stimuli were similar (e.g., interleaving different bird categories) rather than dissimilar (e.g., interleaving bird and painting categories). In Experiment 2, no advantage of interleaving relative to massing stimuli was observed in either the proportion of correct responses or response times on a categorization test for abstract visual stimuli. In Experiment 3 no significant differences between massed and interleaved study conditions were observed on a categorization test for textual materials. Although the results from this study are preliminary, the pattern of results in Experiment 1 suggests that interleaving may be most beneficial when the interleaved stimuli are similar rather than dissimilar.

Table of Contents

List of Figures

Introduction

Throughout the years spent in school, students are continually tested on their memory for conceptual knowledge of information presented in class. Students are expected to spend time in and out of class studying so that they can perform well on tests. What advice can memory researchers give students to make studying as efficient and effective as possible? The research presented here concerns one aspect of this question, namely what is the best way to temporally sequence information when studying in order to promote long-term memory for that information? If a student has two evenings to study for two final exams, for example, she could use that time in different ways. For instance, she could spend the first evening studying the information for one class, and the next evening studying the information for the other class. Another possibility is that she could choose instead to study the information for both classes on both nights. Which method will lead to better performance on the exams? Psychological research on the *spacing effect* and the *interleaving effect* may provide answers to this question.

The Spacing Effect

What is the optimal time interval between learning episodes that will promote long-term retention of information? A large number of studies have shown that spacing out study opportunities over time results in enhanced memory relative to spending the same amount of time in one longer study period. This is called the *spacing effect* or the *distributed practice effect*. The spacing effect has an especially long and voluminous publication record, and there are several reviews of this work (Cepeda, Pashler, Vul, Wixted, & Rohrer, 2006; Dempster, 1988; Dempster, 1989; Donovan & Radosevich, 1999; Dunlosky, Rawson, Marsh, Nathan, & Willingham, 2013; Janiszewski, Noel, &

Sawyer, 2003; Lee & Genovese, 1988; Rohrer, 2009; Rohrer & Pashler, 2007; Rohrer & Pashler, 2010). A 1992 review identified 321 research articles on the topic of distribution and spacing of practice (Bruce & Bahrick, 1992), and the earliest studies pre-date the 20th century (Ebbinghaus, 1885/1913).

The spacing effect has been most often demonstrated with simple materials like word lists or pairs (Bahrick, 1979; Bahrick, Bahrick, Bahrick, & Bahrick, 1993; Bahrick & Hall, 2005; Balota, Duchek, Sergent-Marshall, & Roediger, 2006; Cull, 2000; Finley, Benjamin, Bjork, & Kornell, 2011; Kahana & Howard, 2005; Pyc & Rawson, 2007; Rohrer & Pashler, 2007; Schmidt & Bjork, 1992; Seabrook, Brown, & Solity, 2005, Experiments 1 & 2). In the vast majority of these studies the time interval between spaced presentations of information during encoding is very brief (between a few seconds and a few minutes) and participants are tested on their memory shortly after the information is studied, all within a single experimental session. For instance, Kahana and Howard (2005) had participants study fifteen lists each containing thirty common nouns. The words were presented auditorily at a rate of 1.5 seconds each. In the *massed* condition, each word on the list was repeated three times successively. In the *spaced short* condition, each repetition of a word was separated by between two and six other words from the list. In the *spaced long* condition, each repetition of a word was separated by between six and twenty other words from the list. Participants' memory for each list was measured via a free recall test that occurred about one minute after the list was initially presented. The results showed that free recall was significantly better in the two spaced conditions relative to the massed condition, and significantly better in the spaced long condition relative to the spaced short condition.

The spacing effect has been well established in laboratory research, and its apparent relevance to educational settings has led to a substantial number of studies using more complex and educationally relevant materials as stimuli (Appleton-Knapp, Bjork, & Wickens, 2005; Cepeda, Vul, Rohrer, Wixted, & Pashler, 2008; Cook, 1934; Helsdingen, van Gog, & Merrienboer, 2011; Kornell & Bjork, 2008; Kornell, Castel, Eich, & Bjork, 2010; Le Blanc & Simon, 2008; Rohrer & Taylor, 2006; Rohrer & Taylor, 2007). For instance, Kornell and Bjork (2008) examined the effect of massed versus spaced practice in the task of learning to identify the works of several different artists. Participants studied seventy-two landscape paintings (six each by twelve artists) and were subsequently tested on forty-eight new paintings (four by each artist). In one experiment, participants passively viewed each painting along with the last name of the artist during encoding. Six artists' paintings were studied in a massed fashion and the other six were studied in a spaced fashion. In a massed study block six different paintings by a single artist were presented consecutively, while in a spaced block six different paintings by six different artists were presented. During the test phase, participants viewed a set of paintings that were new to the experiment and decided which of the twelve artists had created each painting. The results showed a clear advantage for spaced presentation, in that the proportion of correct responses was higher for artists who were learned under spaced presentation than for artists learned under massed presentation.

Interestingly, participants were surveyed after the test and asked whether they thought massed or spaced presentation helped them more. More than three quarters of the participants thought massed presentation was as good or better than spaced presentation, even though the results of the experiment clearly contradicted their

intuitions. These results are consistent with a recent study in which participants were given the choice of how to study exemplars from different families of birds (Tauber, Dunlosky, Rawson, Wahlheim, & Jacoby, 2013). Across four experiments a significant majority of participants (at least 78% in each experiment) chose to study the exemplars in a massed rather than spaced manner.

An important implication of the results of Kornell and Bjork (2008) is that spacing effects apply to the transfer of knowledge. It is important to note that in many experimental studies the material that is tested is identical to the material that is studied, so it raises the question of whether or not these effects are limited to the rote memorization of specific material. However, all of the paintings in the final test in Kornell and Bjork (2008) were new to the experiment. Despite not having seen the paintings, participants performed the final test in the spaced conditioned with accuracy just above 60%. This means that the participants were able to generalize what they learned from studying subsets of each artist's paintings and use that knowledge to correctly categorize new material. Helsdingen et al. (2011) also showed transfer of the spacing effect. Participants in their experiment learned skills about how to process case descriptions of crimes, and when they learned in a spaced manner they performed more accurately on transfer tests relating to traffic offenses than participants who learned in a massed manner.

In the interest of exploring the relevance of the spacing effect for educational applications, research on spacing has also been conducted with school-age children and in the classroom (Ambridge, Theakston, Lieven, & Tomasello, 2006; Bloom & Shuell, 1981; Grote, 1995; Rea & Modigliani, 1985; Rea & Modigliani, 1987; Seabrook et al.,

2005, Experiment 3; Solity, 2000). For instance, Bloom and Shuell (1981) investigated the effects of spacing in a high school French classroom. One half of the class was in the spaced group, and that group learned a set of French vocabulary words three times for ten minutes each time on three successive days. The other half of the class was in the massed group, and that group learned the same words during one thirty-minute period on the third day of the study. All students took two cued-recall tests consisting of retrieving the English translations of the French words they learned, once on the third day of the study, and once four days later. The number of words recalled correctly in the immediate test was equal for both spaced and massed groups. However, the number of words recalled on the test given four days later was significantly higher for the spaced group than for the massed group.

It is important to note the fact that the benefits of spaced practice only emerged after a significant delay in this study. The retention interval between study and test varies considerably across studies of the spacing effect, ranging from just a few seconds in many studies to up to five years (e.g., Bahrick et al., 1993). For brief retention intervals, some studies demonstrate an advantage for spaced practice (e.g., Kornell & Bjork, 2008), some studies show no advantage for spaced practice (e.g., Bloom & Shuell, 1981), and still others show an advantage for *massed* practice. For instance, Gagne (1950) created a set of paired associates consisting of an abstract visual form paired with a nonsense syllable. The visual forms fell into one of four categories and the categories were presented in either a massed or spaced manner during study. When tested during the same session on their ability to produce the nonsense syllable when prompted with the visual form, participants in the massed condition showed significantly better performance on the

test than participants in the spaced conditions. The effects of spacing for shorter retention intervals appear to be more varied, but for retention intervals longer than a day the results more consistently support the benefits of spaced practice.

As many of the studies reviewed here vary with respect to the retention interval between study and test, they also vary with respect to the interstudy interval (ISI) or the amount of time or number or stimuli between spaced trials. The ISI can range from just a few seconds (e.g., Kornell & Bjork, 2008) up to several days (e.g., Bahrick et al., 1993). A meta-analysis of 184 studies of the spacing effect (Cepeda et al., 2006) showed that the size of the spacing effect is a function of both the length of the retention interval and the ISI. Generally, spacing effects were found to be stronger as the length of the retention interval and the length of the ISI increased. In addition, while most spaced practice studies use a fixed ISI, some studies have used an expanding ISI. In *expanded practice*, each successive interval increases across the study phase of the experiment. In a study of spaced practice by Landauer and Bjork (1978), participants studied paired associates, first and last names of hypothetical individuals, in three different study conditions. In the massed condition, each paired associate was presented four times in a row. In the uniform spaced condition, each paired associate was separated by a fixed number of other items, e.g., five. In the expanded spaced condition, each paired associate was separated by a number of items that increased linearly across the study phase (e.g., one, two, three). Overall spaced practice resulted in better performance than massed practice on a test given after a thirty-minute retention interval. However, expanded spaced practice also resulted in better performance than uniform spaced practice. While these results were found using simple paired-associate stimuli, expanded practice has also been

demonstrated to be more effective than uniform spaced practice using more complex stimuli (Balota, Duchek, & Logan, 2007; Balota et al., 2006; Cull, Shaughnessy, & Zechmeister, 1996; Morris & Fritz, 2000; Rea & Modigliani, 1985).

In summary, the results of many studies point to the benefit of spaced practice for improving long-term memory. Methods of studying that introduce intervals of time or intervening items between opportunities to study a particular piece of information result in better memory for that information than methods of studying that mass opportunities to study together.

The Interleaving Effect

Is the spacing effect just a matter of inserting time between study opportunities, or is it important that processing of other, intervening information takes place during that time? Research on another study method known as *interleaving* is concerned with the optimal way to organize information during learning episodes in order to promote long-term retention. Several studies have shown that interleaving or mixing different types of information during learning, rather than massing them together, results in improved long-term retention (for reviews, see Dunlosky et al., 2013; Rohrer, 2009; Rohrer & Pashler, 2010; Rohrer, 2012). This is known as the *interleaving effect*. While there has been much less research on interleaving per se than spacing in the memory literature, they are closely related effects. The interleaving effect and the spacing effect both describe a benefit to long-term memory that results from studying information in a manner that is spaced across time. What is distinctive about interleaving is that it focuses not just on the effect of the time that passes between spaced intervals of study, but also on the effect of the other information that is processed during those spaced intervals. Interleaving can be

viewed as an implicit part of most spaced practice studies. For instance, when one set of information is studied multiple times in succession (i.e., massed presentation), the presentations are neither spaced nor interleaved with studying of other information. In contrast, when several different sets of information are studied multiple times in succession (i.e., interleaved presentation), the presentations are both spaced and interleaved with studying of other information. Thus, interleaving and spacing effects are confounded in most studies.

While spacing and interleaving are similar, they are not identical. Generally, interleaving studies are concerned with how mixing different kinds of material while learning can affect long-term memory performance. Spacing studies, on the other hand, are generally focused on the time domain and how much forgetting happens in intervals between learning episodes. Interleaving also seems to have a richer history than spacing in the study of motor skills, and is explored in several studies as "variable" or "mixed" practice (Carson & Wiegand, 1979; Catalano & Kleiner, 1984; Hall, Domingues, & Cavazos, 1994; Keller, Li, Weiss, & Relyea, 2006; Landin, Hebert & Fairweather, 1993; Lee, Magill, & Weeks, 1985; Moxley, 1979; Newell & Shapiro, 1976; Shea & Morgan, 1979; Wrisberg & Ragsdale, 1979). For instance, Landin et al. (1993) had female college students with no basketball experience practice taking set shots over the course of three days. One group of participants practiced taking shots only from twelve feet away from the rim, while the other group practiced taking shots from varying distances of eight, twelve, or fifteen feet from the rim in random order. Three days after the last practice session all participants were tested on shots taken from twelve feet away. The results showed that the variable practice group performed best on the final test, despite the fact

that they had overall taken fewer shots from the distance being tested than the other group.

Like spacing, the interleaving effect has been demonstrated with complex and educationally relevant materials (Birnbaum, Kornell, Bjork, & Bjork, 2013; Carvalho & Goldstone, 2012; Kang & Pashler, 2012; Kornell & Bjork, 2008; Kornell et al., 2010; Le Blanc & Simon, 2008; Richland, Bjork, Finley, & Linn, 2005; Rohrer & Taylor, 2007, Experiment 2; Taylor and Rohrer, 2010; Wahlheim, Dunlosky, & Jacoby, 2011; Zulkiply & Burt, 2013; Zulkiply, McLean, Burt, & Bath, 2012) and in the classroom (Rau, Aleven, & Rummel, 2010; Rohrer, Dedrick, & Burgess, 2014). In a study of interleaving effects with children, Taylor and Rohrer (2010) gave fourth graders four kinds of simple geometry problems to practice, in either massed (learning all of one kind of problem at a time) or interleaved (intermixing different types of problems while learning) conditions. Their study was designed particularly to dissociate the effects of spacing from the effects of interleaving. In the massed condition, students practiced several examples of a single type of geometry problem (e.g., calculating the number of edges of an object) followed by several examples of a different type of problem. However, in between each problem they performed an unrelated filler task, so that practice was spaced over time but not interleaved. In the interleaved condition, students practiced one example of each type of problem in turn, and then at the end of the session performed the same number of filler task trials as the massed group did. So practice in this group was both spaced and interleaved, with total time on task equated between groups.

After practicing the problems, the students took a math test on all four types of problems one day later. In addition to the math test score, the researchers tracked the

progress of the students while practicing the problems the day prior to the test. The results showed that while practicing the problems, the students taking part in massed practice solved more problems correctly than the students taking part in interleaved practice. However, the results of a test taken one day later show the opposite pattern—significantly better test performance for students in the interleaved condition compared with students in the massed condition. Similar to the results of some spacing effect studies, massed practice improved memory performance at a brief retention interval while interleaved practice improved memory performance at a longer retention interval.

One recent study explored interleaving effects in fifth and sixth grade mathematics classrooms in three different schools (Rau et al., 2010). The students practiced fractions problems over five-six days using a computer-based tutoring system, after taking an initial pre-test on their knowledge of fractions. They were divided into four groups depending on how the different types of fractions problems were presented by the tutoring system – in the *massed* condition students were presented with all thirty-six problems of one type consecutively, followed by thirty-six problems of the second type, etc. In the *moderate* condition, the problem type switched after every three problems. In the *interleaved* condition the problem type switched after every single problem. In the *increased* condition, the problem type initially switched after every twelve problems, but this interval gradually decreased so that by the end of the study phase the problem type switched after every single problem. The study phase of the experiment took place during class on five-six consecutive days, with each student assigned to the same condition every day. All the students then took a post-test both one and seven days after completing study.

The results showed significant learning gains from pre- to post-tests only in two of the four study conditions–massed and increased. When the authors median split the participants into a high-prior-knowledge and low-prior knowledge group based on their pre-test scores, they found that the learning gains were only significant for the low prior-knowledge group. Although the authors note that their results are contrary to many other studies that show an advantage for interleaved practice, they suggest that presenting the problems in a massed fashion may facilitate "representational fluency", or the ability of the students to understand and manipulate each type of problem, as opposed to "representational flexibility", or the ability to compare and distinguish one representation from another. Based on their results comparing low and high prior knowledge participants, they speculated that interleaved presentation may only benefit learning when there is a sufficient degree of prior familiarity with the materials to be learned. When prior knowledge is low, they suggest that massed learning may provide necessary scaffolding for students to grasp the basic features of the materials being learned (representational fluency). Once this basic knowledge has been established then interleaving may further facilitate comparisons between types of materials (representational flexibility), as evidenced by the efficacy of the "increased" condition in the above study.

In another recent classroom study, 140 seventh-graders in public school learned math problems in an interleaved or massed manner (Rohrer et al., 2014). The math problems were presented as individual assignments; in the massed condition all problems of one type appeared on a single assignment, while in the interleaved condition problems were distributed across multiple assignments that were given during a nine-week study

period. A brief lecture was given by the teacher with worked examples for each assignment before it was given to the students to complete. Test scores for interleaved material were significantly higher than for massed material (78% versus 38%, respectively) on a final test given in class two weeks after the end of the study period. Unlike the results of Rau et al. (2010), the results of this study indicate that interleaving effects can extend to classroom environments using mathematics problems.

In summary, the interleaving effect is highly related to the spacing effect. Interleaving has been demonstrated to improve memory relative to massing in several studies using different types of materials and settings. While interleaving is being promoted as an effective instructional method in educational settings (e.g., Rohrer, 2012), more classroom-based research is needed to better understand the generalizability of the effect beyond the laboratory.

Main Question

One observation that can be made about many if not all studies of spaced and interleaved practice is that the interleaved stimuli tend to be from the same general semantic category or are highly related. In the study by Kornell and Bjork (2008), for instance, the stimuli were all landscape paintings, and in the study by Taylor and Rohrer (2010), the stimuli were all related geometry problems. It may be of interest to inquire into whether interleaved practice would provide the same benefit to long-term memory when the information to be learned is *not* highly related, or comes from different semantic categories. Thus, the main question of my dissertation can be stated as follows: *is the interleaving effect greater in magnitude when the interleaved information is similar or when the interleaved information is dissimilar?*

One alternative answer to the main question that may be supported by prior research is that the interleaving effect should be greater in magnitude when the information that is interleaved is similar rather than dissimilar. At face value, the results of most interleaving studies to date have primarily provided support for interleaving related information. One proposed theory for why this might occur is called the *discriminative contrast hypothesis* (Birnbaum et al., 2013; Kang & Pashler, 2012; Dunlosky et al., 2013; Rohrer, 2012). By this theory, interleaving highlights the overlapping and non-overlapping information between highly similar categories. Similar categories are thought to have "low contrast," meaning that the differences between them are small and relatively difficult to detect. By highlighting both information that is consistent and information that is variable between categories, interleaving facilitates the process of *induction*, or extracting the invariant features that are diagnostic for a particular category (e.g., one artist's style versus another artist's style). Induction is learning to generalize from memories of relevant prior encounters and can be measured through tasks that require participants to categorize new stimuli that are different from the ones they encountered during study (Kang & Pashler, 2012; Kornell & Bjork, 2008; Kornell, et al., 2010; Wahlheim et al., 2011). It is related to a concept of great interest to the field of education known as *transfer* (Bransford, 1999), which is the ability to flexibly apply knowledge in novel ways and in novel contexts. Both induction and transfer are to be distinguished from specific memory for items, which can be measured through tasks that require participants to indicate whether a specific stimulus was previously encountered during study (Kornell & Bjork, 2008; Wahlheim et al., 2011). Because highly dissimilar categories have "high contrast", this theory would predict that

interleaving should not really benefit discriminating between dissimilar categories, and perhaps would not even provide an advantage over massed learning.

A second alternative answer to the main question is that the interleaving effect should be greater in magnitude when the information that is interleaved is dissimilar rather than similar. There are three proposed explanations for the mechanisms of the interleaving effect that are consistent with this alternative. The first is known as the *attention attenuation hypothesis* (Kornell et al., 2010; Wahlheim et al., 2011). By this theory, the degree to which participants attend to stimuli decreases over the course of a massed block of presentation. It is known that attentional processes are influenced by stimulus-driven biases, one of which is novelty (Reicher, Snyder, & Richards, 1976; Snyder, Blank, & Marsolek, 2008). That is, attention is naturally attuned to novel stimuli in the environment. In a massed block of study each stimulus presented is similar to the previous one, i.e., novelty is low. Therefore the benefit of interleaving over massing may be due to a decrease in attention allocated to stimuli in a massed block. If this is the case, then interleaving dissimilar stimuli should produce a greater degree of novelty across the presentation block than interleaving similar stimuli and better capture participants' attention.

A second theory known as the *retrieval practice hypothesis* (Dunlosky et al., 2013) is also consistent with the alternative that interleaving dissimilar information should be preferable to interleaving similar information. By this theory, the more often information is retrieved from long-term memory into working memory, the greater the likelihood it will be retrieved and/or transferred in the future. In an interleaved block of study of stimuli from *similar* semantic categories, there is a high degree of overlap in

information across each stimulus that is presented. That overlapping information will remain in working memory throughout the study block and not benefit from multiple retrievals. However, in an interleaved block of study of stimuli from *dissimilar* categories there is much less overlapping information. Each time a stimulus is presented participants must retrieve more information from long-term memory into working memory. Thus, alternating between dissimilar categories in an interleaved block of study should result in more retrieval practice and enhance learning overall relative to alternating between similar categories in an interleaved block of study.

A final alternative answer to the main question is that the interleaving effect may be the same in magnitude between similar and dissimilar stimuli. Because the hypotheses put forward in the preceding paragraphs are not necessarily mutually exclusive, the combination of them might produce results that show no significant difference in learning between similar and dissimilar categories of interleaved information. In addition, another theory leads to the prediction of equal magnitude benefits of interleaving similar and dissimilar categories (at least with respect to the current study), namely *transfer-appropriate processing* (TAP) (Morris, Bransford, and Franks, 1977; Roediger, 1990). TAP theory predicts that performance on a test will be highest when the processes invoked during studying are the same as those invoked during the test. By this theory, participants who engage in spaced or interleaved practice at study may perform better at test simply because critical skills needed for the test (i.e., switching between problem types) are practiced while studying. Studying in an interleaved fashion thus better prepares the participant for test conditions in which items are also presented in an interleaved fashion, which is the case for the majority of the interleaving studies

referenced above. With respect to stimulus similarity, the categorization tests used in the present study involved interleaving both similar and dissimilar stimuli to an equal degree at test; thus performance may be equally benefited when studying entails interleaving similar or dissimilar stimuli.

The main question proposed above was investigated in two laboratory-based studies and one classroom-based study. Since the interleaving effect has implications for the design of school curriculum, using educationally relevant materials in two of the studies and exploring the effects of interleaving in the classroom was intended to produce results that may be relevant to educators and the implementation of instructional methods. One of the three studies employed materials that were not very educationally relevant; however, the increased experimental control gained by using low-level stimuli enabled a more rigorous investigation of the role of stimulus similarity in interleaving effects.

Experiment 1

The goal of Experiment 1 was to investigate the role of stimulus similarity in the interleaving effect in a laboratory environment. The experiment consisted of a study phase and a test phase within a single experimental session. Stimuli consisted of photographic images drawn from two distinct groups, paintings and birds. These groups were chosen in part to be the same as stimuli used in prior interleaving studies (i.e., paintings: Kang & Pashler, 2012; Kornell & Bjork, 2008; Kornell et al., 2010; and birds: Wahlheim et al., 2011). This facilitates comparisons between the results of the current study and the results of other published studies. In addition, one of the goals of Experiment 1 was to make inferences from the results that are relevant to educational

settings. Using stimuli from relatively naturalistic categories increases the generalizability and ecological validity of the results of the current study.

The *study condition* variable in Experiment 1 contained three levels: *massed*, *interleaved-similar*, and *interleaved-dissimilar* and was manipulated between subjects. In the *massed* condition, each block of study contained stimuli from a single category of paintings or birds (e.g., all Seurat or all Chickadee). In the *interleaved-similar* condition, each block of study contained stimuli from multiple categories within either paintings or birds (e.g., Chickadee, Sparrow, Warbler...). Finally, in the *interleaved-dissimilar* condition, each block of study contained stimuli from multiple categories across paintings and birds (e.g., Seurat, Chickadee, Pessani, Sparrow…). After the study phase, all participants were tested on their classification performance for novel stimuli to examine how the process of induction or transfer was influenced by the relatedness of the stimuli that were interleaved.

Participants

Ninety-seven participants (69 female; age: $M = 20.7$, $SD = 5.2$) from the University of Minnesota were recruited to participate for course credit or a Target gift card. The number of participants was determined by an examination of effect sizes in prior interleaving studies. In the Kornell and Bjork (2008) study, for example, the reported Cohen's *d* measures comparing massed versus interleaved practice were 0.99 and 1.28 for Experiments 1a and 1b, respectively. In the Kang and Pashler (2012) study, the reported Cohen's *d* measures for the same comparison were 0.78 and 0.56 for Experiment 1 and Experiment 2, respectively. These two studies were very similar to Experiment 1 in that they compared interleaved versus massed practice using paintings or

birds as stimuli. Using Cohen's (1988) power table, assuming a power level of 0.90 and an effect size of 0.90 (computed as a mean of the four effect sizes given above), the number of participants per group needed to find a significant effect at alpha=0.05 is twenty-eight. This is roughly consistent with the between-subjects *N*'s used in the same referenced studies, which were thirty-six per group in Experiment 1b of Kornell and Bjork (2008), twenty-two per group in Experiment 1 and thirty in Experiment 2 of Kang and Pashler (2012). Between thirty-one and thirty-three participants were recruited per group in the current study. Two participants were discarded from the analyses due to errors in complying with the experiment procedure. Ten additional participants were excluded from the analyses because their mean test scores were not significantly above chance performance. Thus, the total *N* was 85–with 27, 30, and 28 participants each in the massed, interleaved-similar, and interleaved-dissimilar groups, respectively.

Design and Materials

The experiment used a one-way between-subjects design. The independent variable *study condition* had three levels (*massed, interleaved-similar, and interleaved-dissimilar*). The dependent variables were the proportion of novel stimuli correctly categorized at test and the mean response time for test items scored as correct.

Stimuli from the birds group were chosen according to their biological classification by taxonomic order, *passeriformes* (perching birds). The twelve bird families used in the current experiment were identical to the twelve bird families used in Wahlheim et al. (2011). The stimuli used by Wahlheim et al. were procured for use in this experiment in order to facilitate a direct comparison of results. Stimuli from the paintings group were chosen according to their classification by landscape artist. The

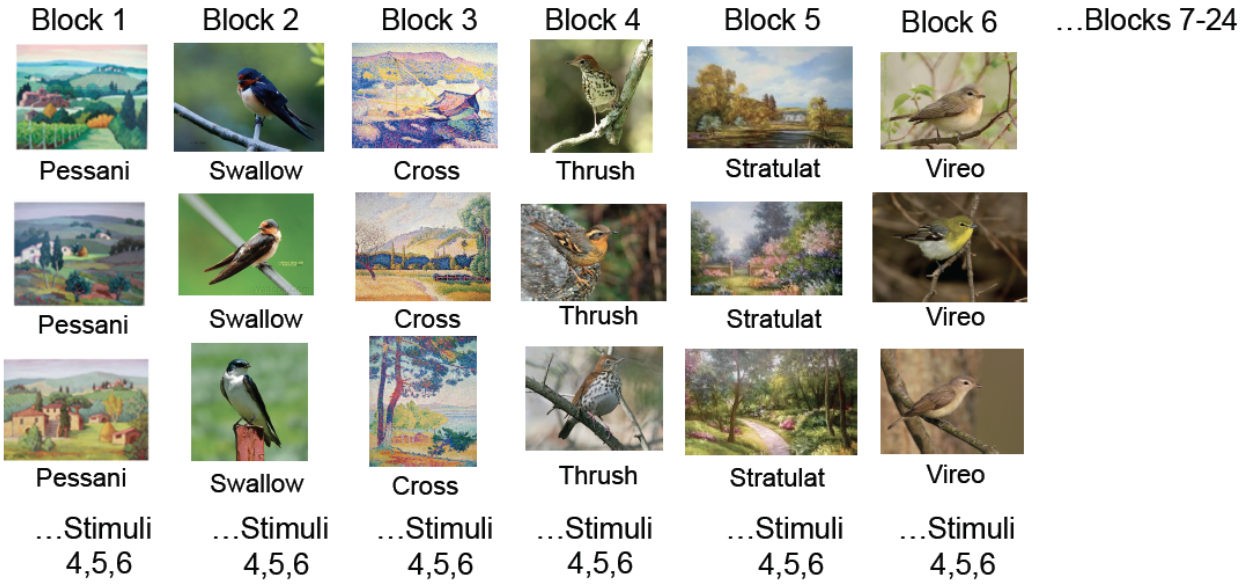
same twelve landscape artists used by Kornell & Bjork (2008) and Kornell et al. (2010) served as the twelve unique categories of those stimuli. The stimuli used by Kornell et al. were also procured for use in this experiment in order to facilitate a direct comparison of results.

Similarly following the methods of the studies referenced above, ten exemplars were chosen from each of the twenty-four categories for a total of 240 unique stimuli, with six of the ten exemplars from each category serving as study stimuli and the remaining four serving as novel stimuli to test participants' classification performance. This resulted in a total of 144 unique stimuli presented during study and ninety-six unique stimuli presented at test. Participants in each of the three study conditions were presented with all twenty-four unique categories of birds and paintings during a single-session experiment.

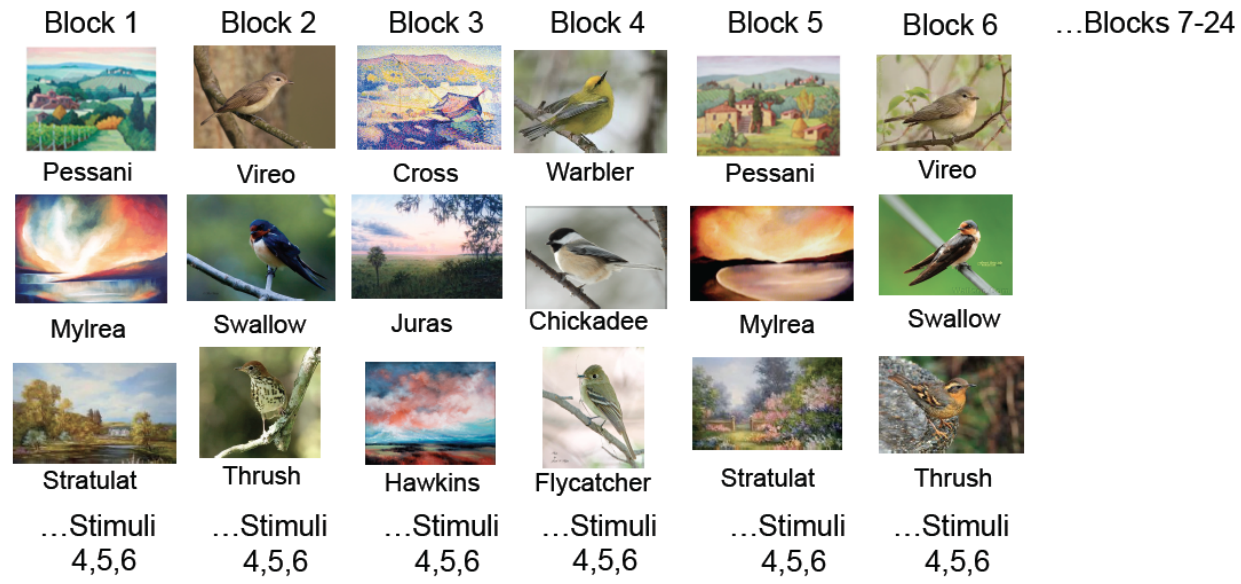
The different patterns of stimulus presentation for each of the three groups are depicted in Figure 1. Participants who were in the *massed* study condition studied all twenty-four categories of stimuli one at a time in a massed manner (i.e., each block in the study phase consisted of either six paintings by a single artist or six exemplars from a single bird family). Participants in the *interleaved-similar* study condition studied all twenty-four categories of stimuli in an interleaved manner within each group (i.e., each block in the study phase consisted of six paintings, each by a different artist, or six birds, each from a different family). Participants in the *interleaved-dissimilar* study condition studied all stimuli from both groups in an interleaved manner (i.e., each block in the study phase consisted of three birds from three different families and three paintings by three different artists, and each trial alternated between a bird and a painting).

Figure 1

Massed study condition



Interleaved-similar study condition



Interleaved-dissimilar study condition

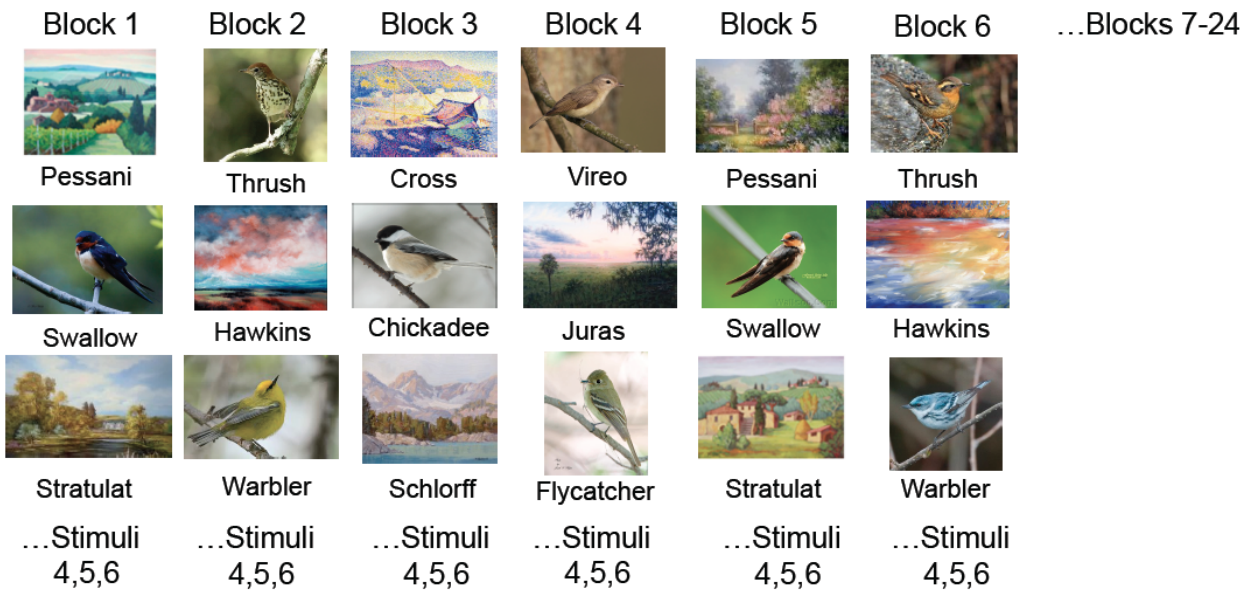


Figure 1. Depiction of the three between-subjects study conditions for Experiment 1.

The study phase consisted of twenty-four blocks each containing six stimuli. In the massed condition each block consisted of presentations of six stimuli from one category. In the interleaved-similar condition, each block contained one exemplar each from six different bird families, or one exemplar each from six different artists. Each set of six artists or six bird families presented in a block remained consistent throughout the study phase (e.g., bird families one through six were always presented together in the same block, and bird families seven through twelve were always presented together in the same block). The same categories were paired together across participants as well. This design ensured the same number of presentations (six) of each stimulus category during the study phase and the same number of comparisons (five) to other unique categories of stimuli within an interleaved block. In the interleaved-dissimilar condition, each block contained one exemplar each from three different bird families and one exemplar each from three different artists. Again, each set of three artists and three birds presented in a block remained consistent throughout the study phase and across participants. The ISI for each presentation of an exemplar from a particular stimulus category in the study phase was the same in both interleaved-similar and interleaved-dissimilar conditions, twenty-three trials. The ISI always consisted of one trial per each of the other stimulus categories.

The test phase consisted of ninety-six items, four from each of the twelve categories of paintings and twelve categories of birds. The order of presentation in the test phase was the same for all participants, and pseudo-randomized so that no more than four birds or paintings were presented in consecutive trials, and no single bird family or artist was repeated in consecutive trials. All stimuli in the study and test phases were

presented in color and sized to fit into a 15 x 15 centimeter square on a fifteen-inch monitor display. The images varied somewhat in size and dimension but subtended approximately seventeen degrees of visual angle.

Procedure

Participants were instructed prior to the beginning of the experiment that they should study each stimulus and try to learn the name associated with it. They were also informed that they would be asked to categorize a new set of stimuli from the same categories at the end of the study phase. Participants were seated in front of a computer at a distance of roughly fifty centimeters. During each study trial, a stimulus was presented on the computer screen for six seconds with the name of the stimulus (e.g., Swallow, or Pessani) appearing below the image. During each trial, participants were instructed to speak aloud the name of the stimulus and study it while it appeared on the screen. After all 144 stimuli were presented participants were given a set of paper mazes to complete and were timed for three minutes before proceeding to the test phase.

During each test trial, a new stimulus was presented on the computer screen with the names of the twelve possible categories (birds or artists) to which it might belong appearing below the image. Each of the twelve category names was presented along with a letter (a-l) corresponding to a letter key on the keyboard. Participants were instructed to choose via keyboard press the category name that correctly identified the stimulus. The stimulus image remained on the screen until the participant responded. The next trial began one second after the participant's response. No feedback was given during test trials to avoid learning effects that may otherwise have altered performance from the beginning to the end of the test phase. Following the test phase a brief survey was

administered to assess participants' prior knowledge of the materials learned in the study and their perception of the difficulty of the task.

Predictions

See Figures 2-4 for graphical depictions of the predictions from both alternative answers to the main question. On the one hand, if the discriminative contrast hypothesis is correct, the proportion correct of new stimuli categorized should be significantly higher in the interleaved-similar group than either the massed or interleaved-dissimilar groups. On the other hand, if the attention attenuation and/or retrieval hypotheses are correct, then the proportion correct of new stimuli categorized should be significantly higher in the interleaved-dissimilar group than either the massed or interleaved-similar groups. Based on the results of many prior interleaving studies, a benefit of interleaved (whether of dissimilar or similar categories) over massed presentation was predicted.

Figure 2

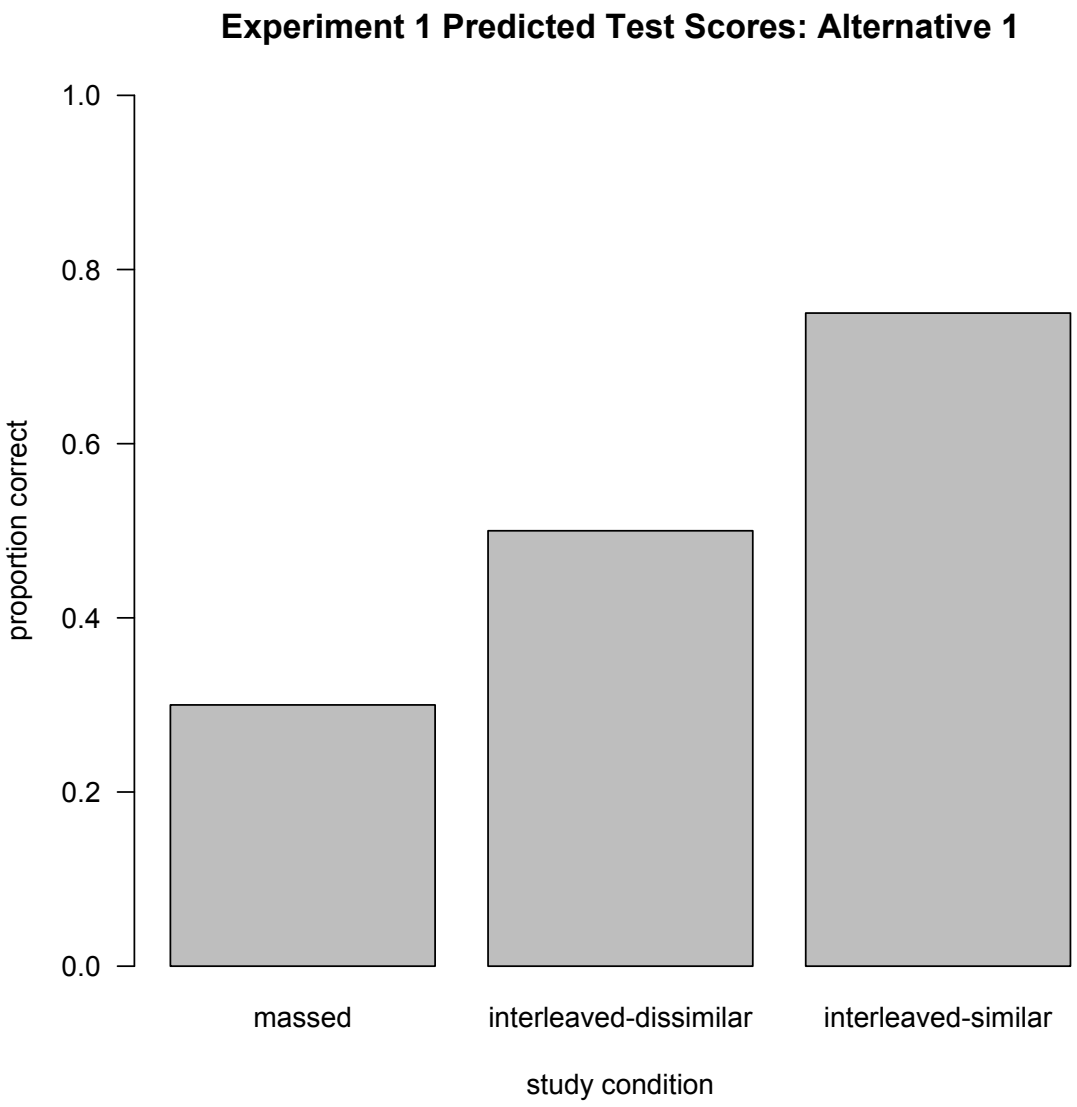


Figure 2. Bar graphs depicting predicted results for Experiment 1 based on the hypothesis that interleaving similar materials should result in better categorization test performance than interleaving dissimilar materials.

Figure 3

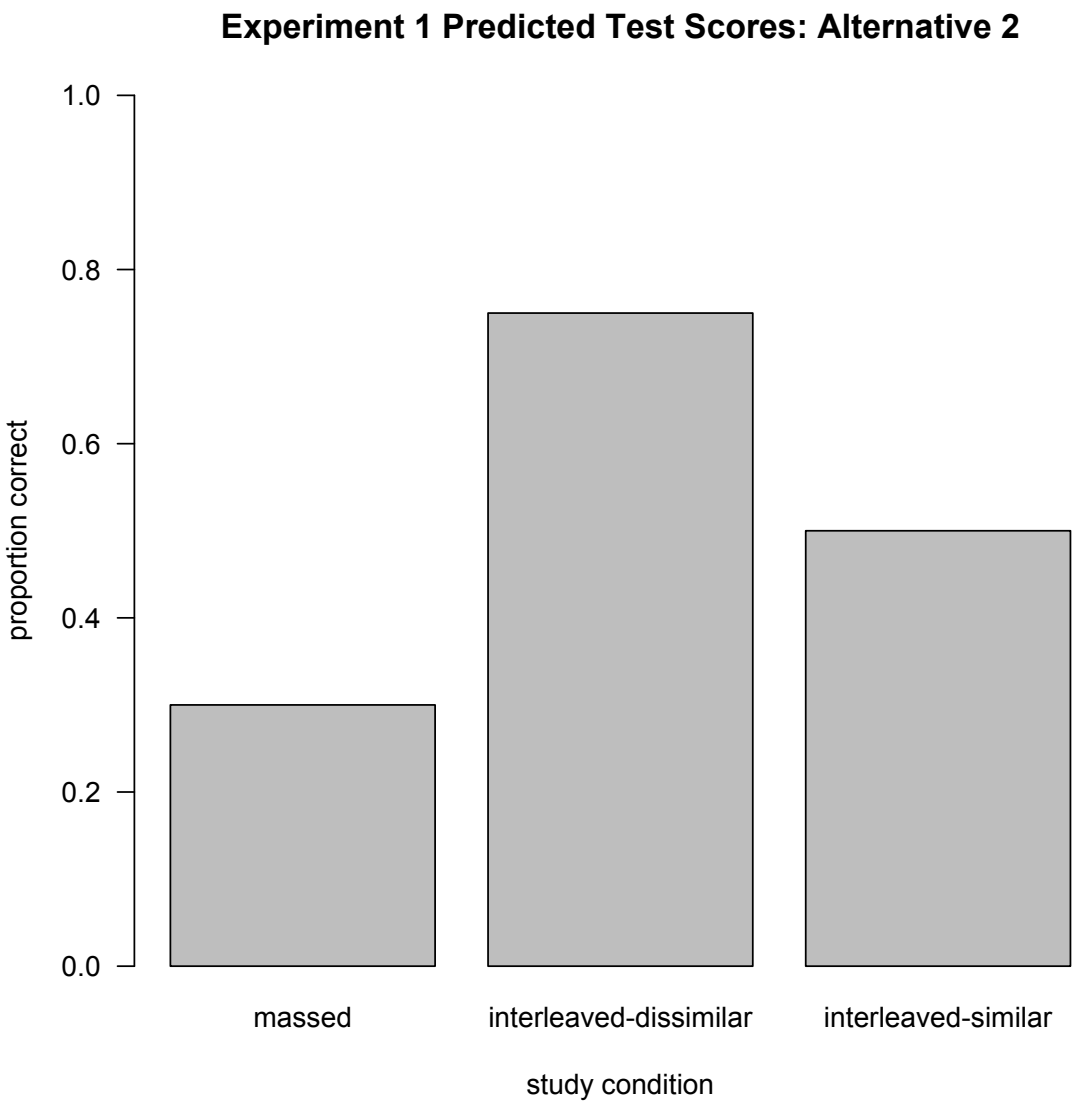


Figure 3. Bar graphs depicting predicted results for Experiment 1 based on the hypothesis that interleaving dissimilar materials should result in better categorization test performance than interleaving similar materials.

Figure 4

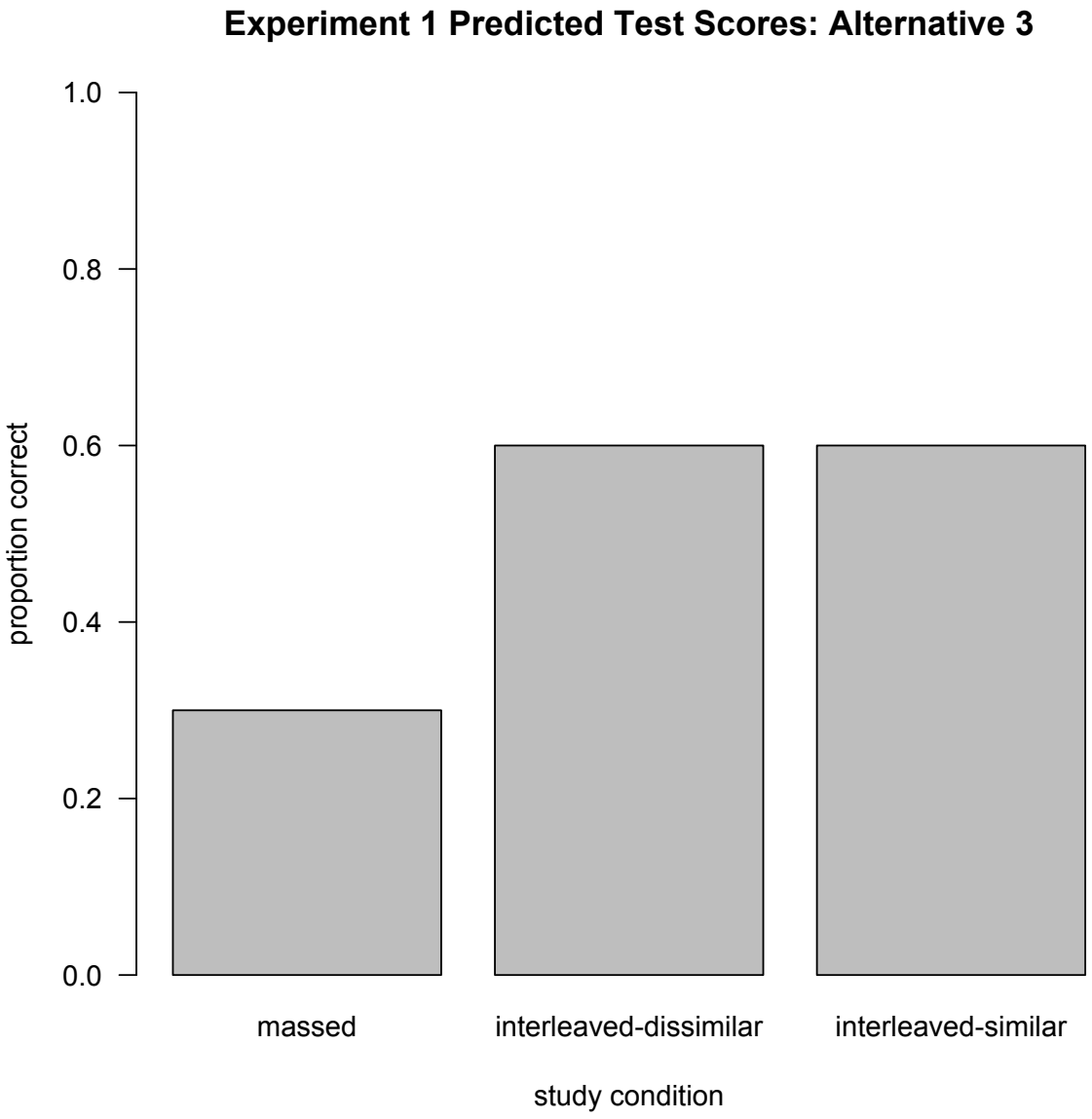


Figure 4. Bar graphs depicting predicted results for Experiment 1 based on the hypothesis that interleaving either similar or dissimilar materials should result in the same level of categorization test performance.

Results

Proportion correct. Figure 5 shows a summary of the categorization test results for the proportion of correct responses. The mean proportion of correct responses across all three groups was rather low (M = .345, SD = .123), but it was significantly higher than chance performance, (H_o = .083): $t(84) = 19.69$, $p < .001$.

A one-way ANOVA with study condition as a between-subjects factor was conducted to compare the effect of study condition on the proportion of correct responses in massed, interleaved-similar, and interleaved-dissimilar conditions. There was a significant effect of study condition on the proportion of correct responses on the test at the $p < .05$ level, $F(2, 82) = 22.06$, $MS_e = .010$, $p < .001$, $\eta_p^2 = .35$. Follow up t-tests indicated the mean test scores for the massed condition (M = .257, SD = .083) were significantly lower than the mean test scores for both the interleaved-similar condition (M = .433, SD = .127), $t(55) = 6.11$, $p < .001$ (Bonferroni corrected $p < .001$), $d = 1.65$, and the interleaved-dissimilar condition (M = .337, SD = .080), $t(53) = 3.66$, $p < .001$ (Bonferroni corrected $p = .002$), $d = 1.01$. Furthermore, the mean test scores for the interleaved-similar condition were significantly higher than for the interleaved-dissimilar condition, $t(56) = 3.39$, $p = .001$ (Bonferroni corrected $p = .004$), $d = .91$.

Figure 5

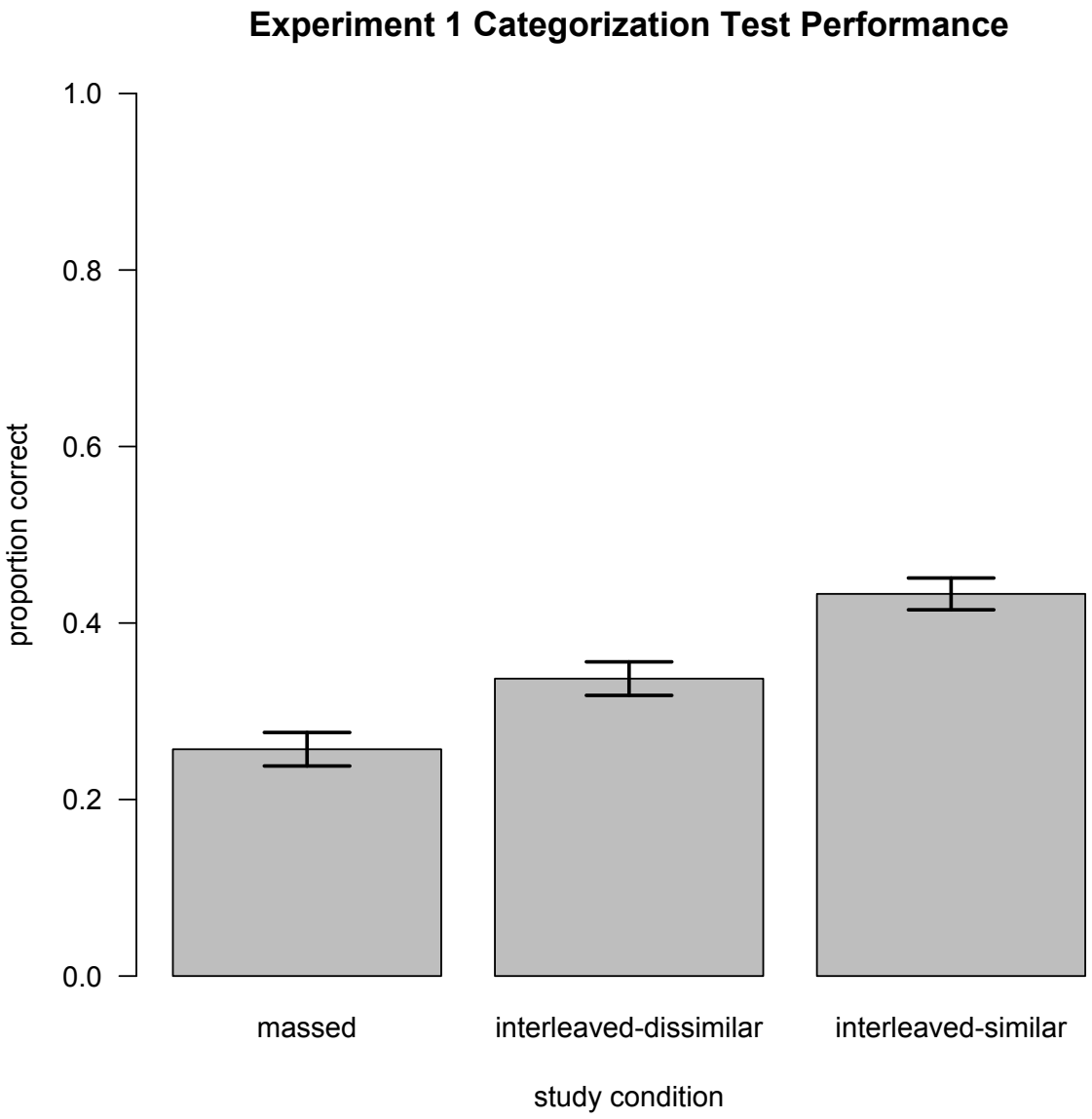


Figure 5. Bar graph depicting the mean proportion of correct responses on the 96-item categorization test for Experiment 1 (N=85). Error bars represent the standard error of the mean. All pairwise mean comparisons are significant at the $p < .05$ level.

Response times. Figure 6 shows a summary of the categorization test results for response times in milliseconds for trials with correct responses. Response times that fell above 2.5 standard deviations of a participant's mean response time or were below 200 milliseconds were discarded from the analyses. This resulted in the omission of 3.2% of total trials. The mean response time across all three groups was 5412 (SD = 1671). There was a significant effect of study condition on response time at the $p < .05$ level, $F(2, 92) = 3.92$, $MS_e = 2{,}611{,}712$, $p = .024$, $\eta_p^2 = .09$. Follow up t-tests indicated that mean response times for the massed condition (M = 6079, SD = 2178) were not significantly different than those for the interleaved-similar condition (M = 5316, SD = 1487), $t(55) = 1.56$, $p = .125$. However, the mean response times for the massed condition were significantly different than those for the interleaved-dissimilar condition (M = 4872, SD = 995), $t(53) = 2.66$, $p = .010$ (Bonferroni corrected $p = .031$), $d = .73$. Furthermore, the mean response times for the interleaved-similar condition were not significantly different from those for the interleaved-dissimilar condition, $t(56) = 1.33$, $p = .190$. Most important, the response time results indicate that there were no tradeoffs between speed of response and the accuracy effects.

Figure 6

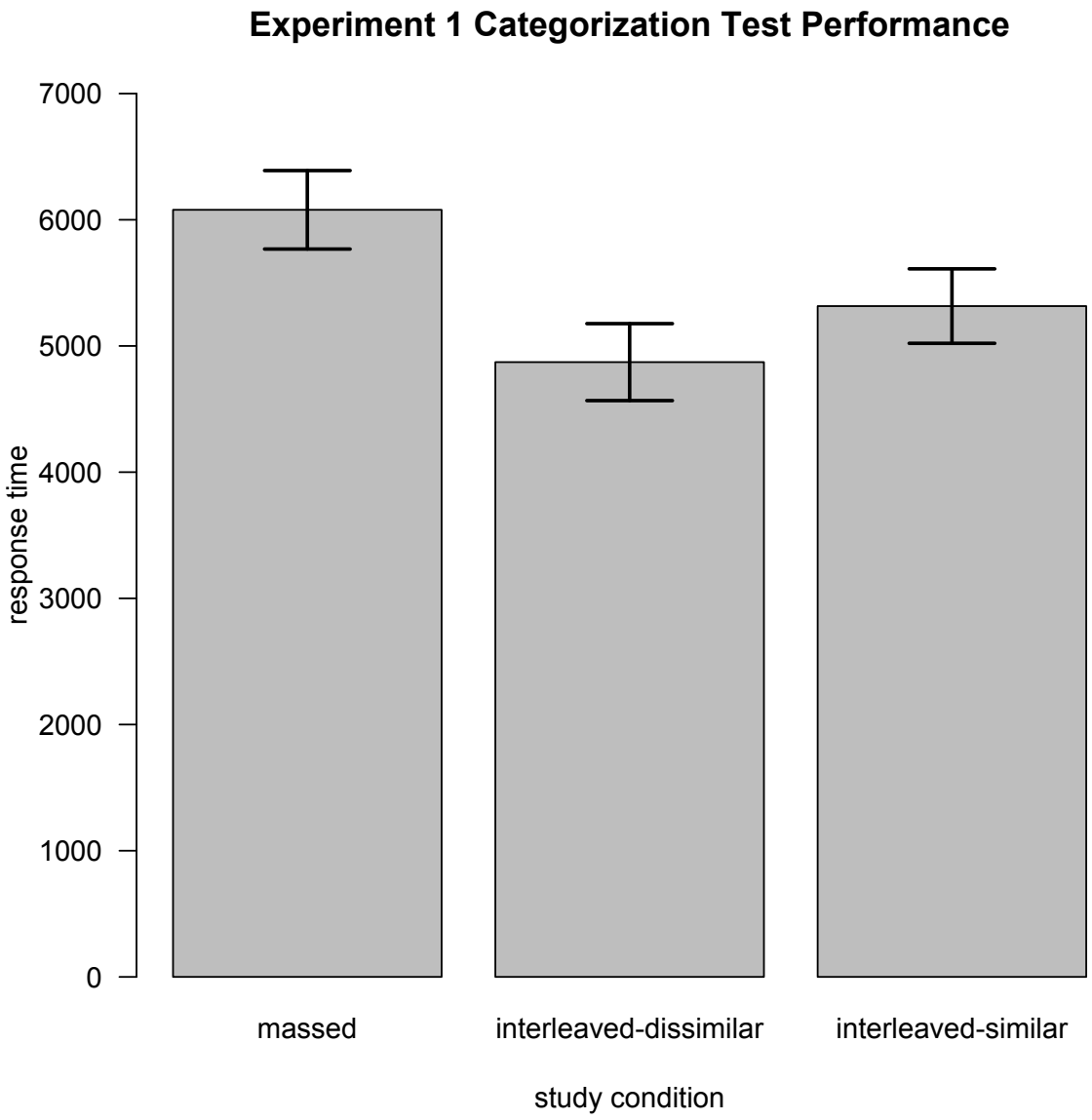


Figure 6. Bar graph depicting the mean response times for correct trials on the 96-item categorization test for Experiment 1. Error bars represent the standard error of the mean. The pairwise mean comparison between the massed and interleaved-dissimilar conditions is significant at the $p < .05$ level. No other pairwise comparisons are significant.

Post-experiment survey. After completing the experiment, participants were given a three-item survey. On one item they were instructed to indicate on a scale of one to seven how difficult they found the task, with one labeled "not difficult" and seven labeled "very difficult". The mean rating was 5.36 (*SD* = .99). This indicated that participants found the task rather difficult, which is in line with their relatively low test scores. There was also a significant negative correlation between the difficulty ratings and proportion correct test scores, $r(83) = -.24$, $p < .05$, which suggests that participants had some meta-cognitive awareness of their own level of performance on the test. The mean difficulty rating did not differ significantly across the three study conditions, however (ts < 1).

The other two items on the survey inquired into participants' prior knowledge of the stimulus categories used in the experiment. Participants were instructed to indicate their level of familiarity with the birds and paintings on a three-point scale: "unfamiliar", "somewhat familiar", and "very familiar". For the birds, 70.5% of participants indicated that they were unfamiliar with the stimuli, while the remaining 29.5% indicated that they were somewhat familiar (no participants chose the "very familiar" option). For the paintings, 85.8% of participants indicated that they were unfamiliar with the stimuli, 11.8% indicated that they were somewhat familiar, and 2.4% indicated that they were very familiar. In order to investigate the possible impact of familiarity or prior knowledge on the interleaving effects, the data were divided into two groups, one group of participants who responded "unfamiliar" to both items ($N = 53$), and one group of participants who responded "somewhat familiar" or "very familiar" to at least one of the items ($N = 32$). A 2 (familiarity: low, high) x 3 (study condition: massed, interleaved-

similar, interleaved-dissimilar) between-subjects ANOVA was used to analyze the proportion of correct responses on the categorization test. The main effect of familiarity on test score was not significant, $F(1, 79) = .00$, $MS_e = .010$, $p = .943$. The interaction between study condition and familiarity was also not significant, $F(2, 79) = .39$, $MS_e = .010$, $p = .676$. This shows that *a priori* familiarity with the categories did not impact (i.e., was not confounded with) learning.

Discussion

The results of Experiment 1 replicated prior studies using the same stimuli that demonstrated an advantage of studying in an interleaved fashion over studying in a massed fashion when tested at a brief retention interval. The mean test scores for both interleaved-similar and interleaved-dissimilar groups were significantly higher than the mean test scores for the massed group. The significant difference in response times between the interleaved-dissimilar group and the massed group was also consistent with the pattern of the test scores. The mean response times for this experiment should be interpreted with some caution, however, due to the complexity of making a response on the keyboard to twelve different response alternatives. Response time data has not previously been reported in published studies using these stimuli and significant effects have been reported for test scores only. The combined results of the proportion correct and response time data still suggest an advantage of interleaved over massed practice.

The significantly higher test scores for the interleaved-similar group relative to the interleaved-dissimilar group were in line with one alternative hypothesis to the main question of the current study. Namely, interleaving similar information was more advantageous for learning than interleaving dissimilar information. The results here

support the discriminative contrast explanation, which states that the benefit of interleaved practice results from facilitating comparisons between items from different, low-contrast categories (i.e., birds vs. birds or artists vs. artists). Because the interleaved categories in the interleaved-dissimilar condition were high- rather than low-contrast (i.e., birds vs. artists), these comparisons may not have been as helpful as in the interleaved-similar condition. However, the significantly higher test scores in the interleaved-dissimilar condition relative to the massed condition suggest that even interleaving high-contrast categories is superior to massing them during study.

Taken together, the proportion correct and response time results from Experiment 1 show that interleaving birds and paintings during study produces better performance on the categorization test compared with massing them during study, but that the advantage of interleaving is largest when the interleaved information is similar rather than dissimilar.

Experiment 2

In Experiment 1, the similarities and dissimilarities between different categories of stimuli had face validity, but they were not readily quantified. In Experiment 2, artificial categories were created and used as stimuli, so that the magnitudes of the similarities between different interleaved stimuli were strictly quantified and well controlled. Artificial categories in the visual domain have been used in some prior studies of spacing and interleaving (Carvalho & Goldstone, 2012; Gagne, 1950; Kurtz & Hovland, 1956; Zulkiply & Burt, 2013). The results of these studies have been somewhat mixed. As mentioned previously, the results of Gagne (1950) showed a learning advantage for massed presentation for learning artificial categories; however, the results

of Kurtz & Hovland (1956) showed an advantage for interleaved presentation. More recently, Carvalho and Goldstone (2012) and Zulkiply and Burt (2013) both found significant interleaving effects for abstract visual stimuli, but only when the stimulus categories were highly similar or of low discriminability – results that support the discriminative contrast hypothesis discussed previously. In the current study, artificial visual categories were created as a useful way to control the degree of similarity between categories. The main question was again whether the interleaving effect would be greater in magnitude when the interleaved information is similar or when the interleaved information is dissimilar.

The categories used were "plaids," which are grayscale images composed of overlapping Gabor patches of varying orientations and spatial frequencies (for examples see Figure 7). The similarities between plaid categories can be manipulated, and such categories have proven useful in recent learning and memory experiments conducted in the Marsolek lab. In the figure, "LLH" refers to a plaid consisting of a low-frequency Gabor patch in the horizontal orientation, a low-frequency Gabor patch in the vertical orientation, and a high-frequency Gabor patch in the diagonal orientation.

Participants

Sixty-four undergraduate participants (49 female, age: $M = 19.7$, $SD = 2.2$) from the University of Minnesota were recruited to participate for course credit or a Target gift card. Because the materials used in the current experiment were abstract visual stimuli rather than birds or paintings, the number of participants recruited was based in part on the reported *N*'s in Gagne (1950) and Kurtz and Hovland (1956) (the more recent studies using artificial categories were not known by the author at the time the current study was

designed). While it is not possible to estimate Cohen's *d* from the results reported in either study, significant effects were found (albeit a massed advantage) in Gagne (1950) with four categories of stimuli and fifteen subjects per group. In Kurtz and Hovland (1956), significant interleaving effects were found with four categories of stimuli and thirteen subjects per group. This suggested that with eight categories of stimuli, 64 participants in a within-subjects experiment should provide sufficient power. Two participants were excluded from the analyses because their mean test scores were not significantly above chance performance, resulting in a total *N* of 62 for the analyses.

Figure 7

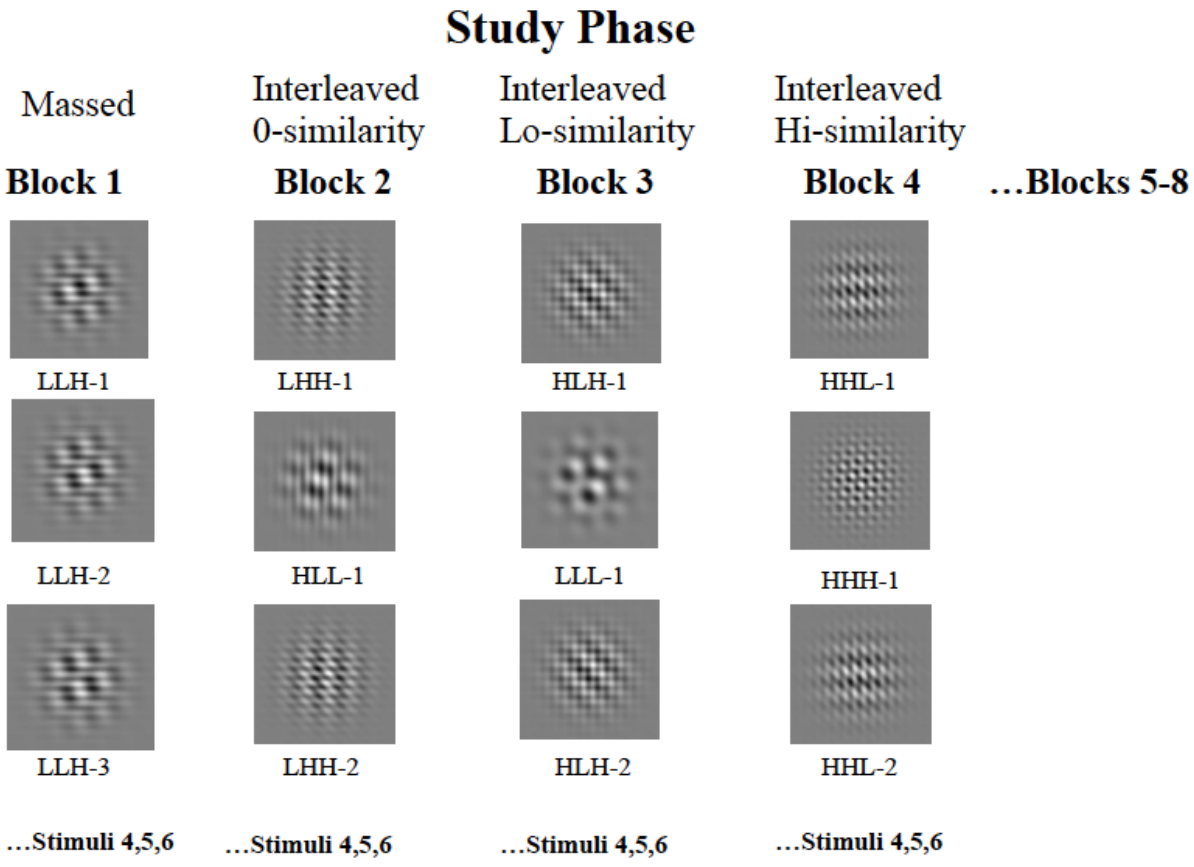


Figure 7. Depiction of the four within-subjects study conditions for Experiment 2.

Design and Materials

Unlike Experiment 1, this experiment utilized a within-subjects design and the independent variable *study condition* had four instead of three levels (*massed, interleaved-0 similarity, interleaved-low similarity, and interleaved-high similarity*). Like Experiment 1, the dependent variables were the proportion of novel stimuli correctly categorized at test and the mean response time for test items scored as correct.

The stimuli consisted of eight unique categories of plaids made up of different combinations of two spatial frequencies (high versus low) at three different orientations (horizontal, vertical, and diagonal). Thus, the eight categories were HHH, LLL, HHL, HLL, HLH, LHL, LLH, and LHH. Each category was given a particular letter to designate its category name. The letters chosen were S, T, U, V, W, X, Y, and Z. This letter designation remained constant for all participants (e.g., category HHH was always designated with the letter U).

Stimulus similarity was defined as the number of shared features between each pair of categories. For example, the stimuli HHH and LLL served as a "0-similarity" pair of stimuli, in that none of the three features are shared between them. For other examples, the stimuli LHH and HLH served as "low-similarity" stimuli in that they share one of three features, and the stimuli LHH and LHL served as "high-similarity" stimuli, in that they share two of three features. This enabled a parametric design in which similarity had three equidistant levels ranging from zero to low to high. Twenty exemplars from each of the eight categories were used in the experiment, ten in the study phase and ten novel stimuli to be classified during the test phase. The exemplars within each category differed only in the phases of their Gabor patches.

The study design is depicted in Figure 7. During a *massed* study block, participants studied ten exemplars from a single category (e.g., LLH). During an *interleaved-0 similarity* study block, participants also studied ten exemplars, five each from two 0-similarity categories that were presented in an interleaved manner (e.g., LHH and HLL). In an *interleaved-low similarity* study block, participants studied five exemplars each from two low-similarity categories in an interleaved manner (e.g. HLH and LLL). Finally, in an *interleaved-high similarity* study block, participants studied five exemplars each from two high-similarity categories (e.g., HHL and HHH). Thus for any given subject, two categories each were presented in the massed, interleaved-0 similarity, interleaved-low similarity, and interleaved-high similarity conditions. The study phase consisted of eight blocks, two per study condition. The ISI between the two blocks of a particular study condition was always thirty trials, and consisted of one block of trials per each of the other three study conditions.

In order to avoid item effects and ensure that each of the eight stimulus categories was represented in each of the four study conditions an equal number of times across the participant sample, four unique experiment scripts were created. Within each of these four scripts, four additional versions were created in order to ensure that each of the study conditions was represented an equal number of times at the beginning and end of the study phase. This prevented primacy and recency effects from contaminating the interpretation of the results. A total of sixteen unique versions of the study phase were thus used. All participants took the same categorization test, which consisted of eighty items, ten new stimuli each from the eight stimulus categories. The order of presentation in the test phase was the same for all participants and was pseudo-randomized so that no

more than three stimuli from a category were presented in consecutive trials. All stimuli were 15 x 15 centimeters in size, subtending approximately seventeen degrees of visual angle on a fifteen-inch monitor display.

Procedure

Participants were instructed prior to the beginning of the experiment that they should study each stimulus and try to learn the letter associated with it. They were also informed that they would be asked to categorize a new set of stimuli from the same categories at the end of the study phase. Participants were seated in front of a computer at a distance of roughly fifty centimeters. For each study trial, a plaid stimulus was presented for three seconds along with the name of the category (e.g., Category W) appearing below the image. After the stimulus disappeared from view, "Which category?" appeared on the screen. The participants were instructed to make a button press to indicate which category the stimulus that was just presented belonged to, which was done to encourage category induction and sufficient encoding of the stimuli. After they made the button press the next trial began.

The results of initial piloting of participants on this procedure showed that performance was close to floor (not significantly above chance at .125), indicating that categorization of these plaid stimuli was rather difficult. Because of this it was decided to add an additional study phase to the experiment, so that all participants viewed each stimulus twice in the same order they were originally presented. The procedure in the second study phase was identical to the first. Upon completing the second study phase, participants were given a set of paper mazes and timed for three minutes before proceeding to the categorization test.

For each test trial, a new stimulus was presented on the computer screen, with the letters designating each of the eight possible categories to which it might belong appearing below the image. Participants were instructed to choose via keyboard press the category letter that correctly identified the stimulus. The stimulus image remained on the screen until the participant responded. The next trial began one second after the participant's response. No feedback was given during test trials to avoid learning effects that may have altered performance from the beginning to the end of the test phase. Following the test phase a brief survey was administered to assess participants' prior knowledge of the materials learned in the study and their perception of the difficulty of the task.

Predictions

See Figures 8-10 for graphical depictions of the predictions from both alternative answers to the main question. Similar to the predictions for Experiment 1, if the discriminative contrast hypothesis is correct the proportion correct of new stimuli should be highest in the interleaved-high similarity group, slightly lower in the interleaved-low similarity group, and even lower in the interleaved-no similarity group. All three interleaved conditions are predicted to produce better classification performance than the massed condition.

On the other hand, if the attention attenuation and/or retrieval hypotheses are correct, then the proportion correct of new stimuli should be highest in the interleaved-0 similarity group, slightly lower in the interleaved-low similarity group, and even lower in the interleaved-high similarity group. Again, all three interleaved conditions are predicted to produce better classification performance than the massed condition.

Figure 8

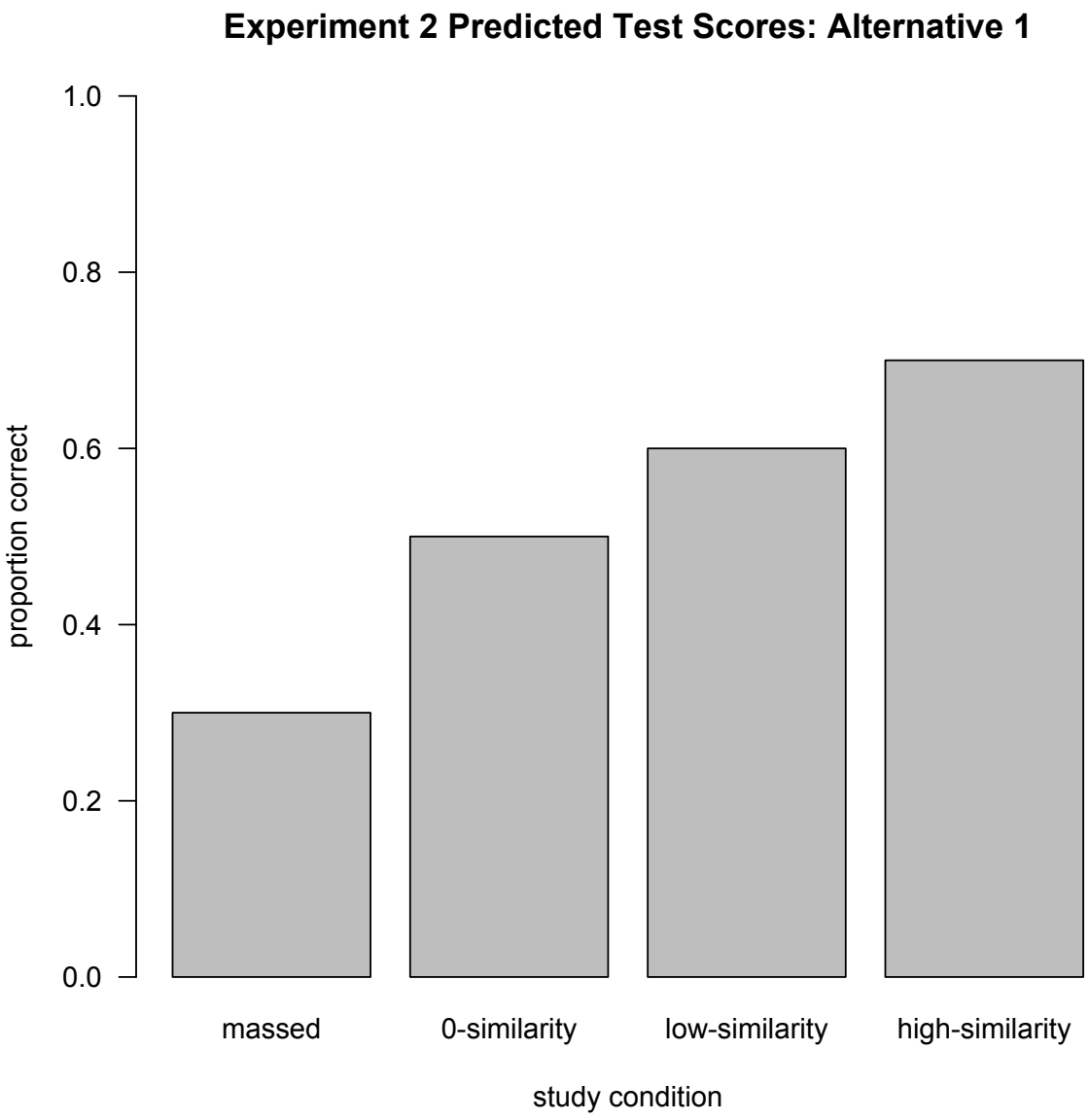


Figure 8. Bar graphs depicting predicted results for Experiment 2 based on the hypothesis that interleaving similar materials should result in better categorization test performance than interleaving dissimilar materials.

Figure 9

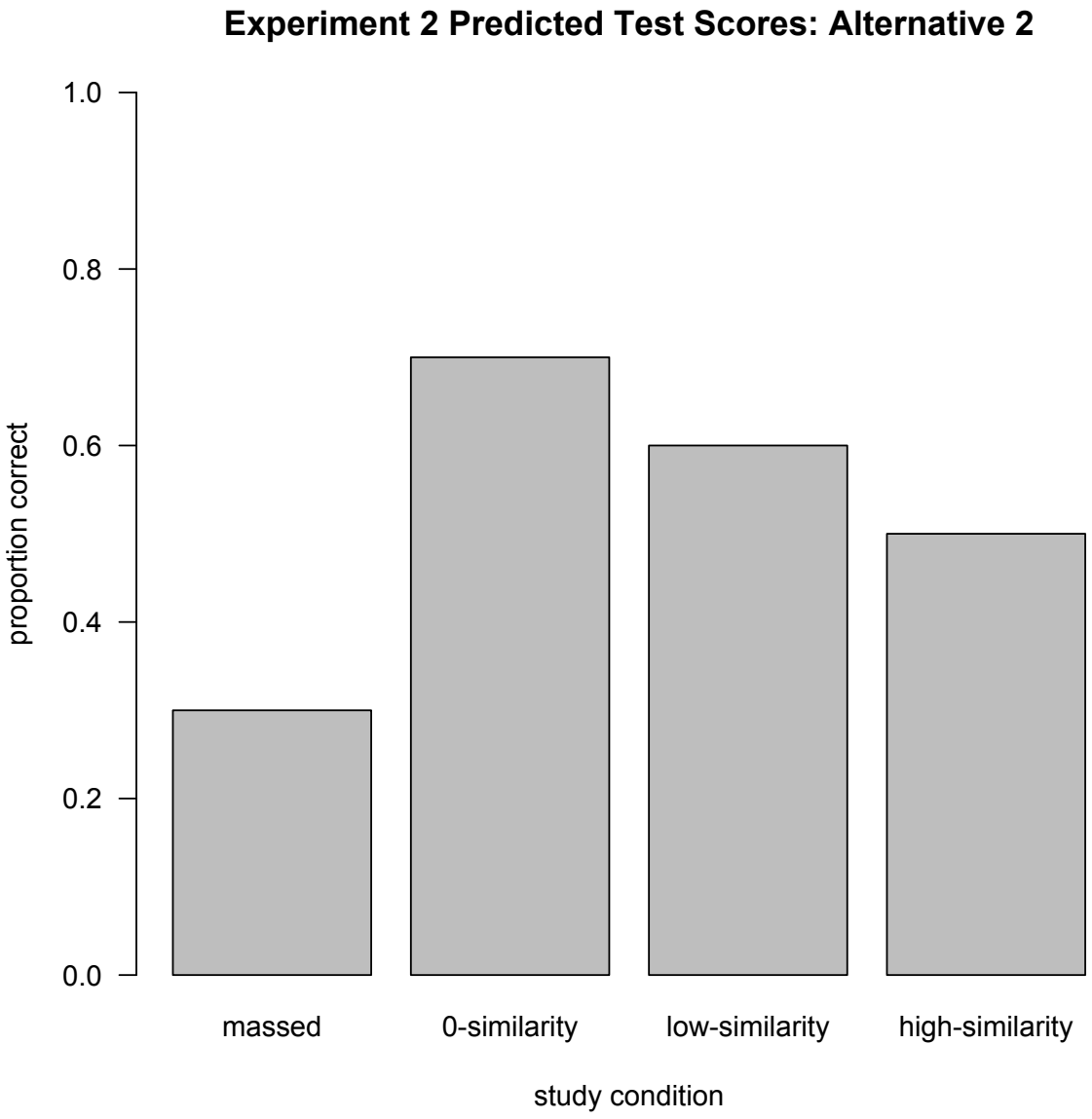


Figure 9. Bar graphs depicting predicted results for Experiment 2 based on the hypothesis that interleaving dissimilar materials should result in better categorization test performance than interleaving similar materials.

Figure 10

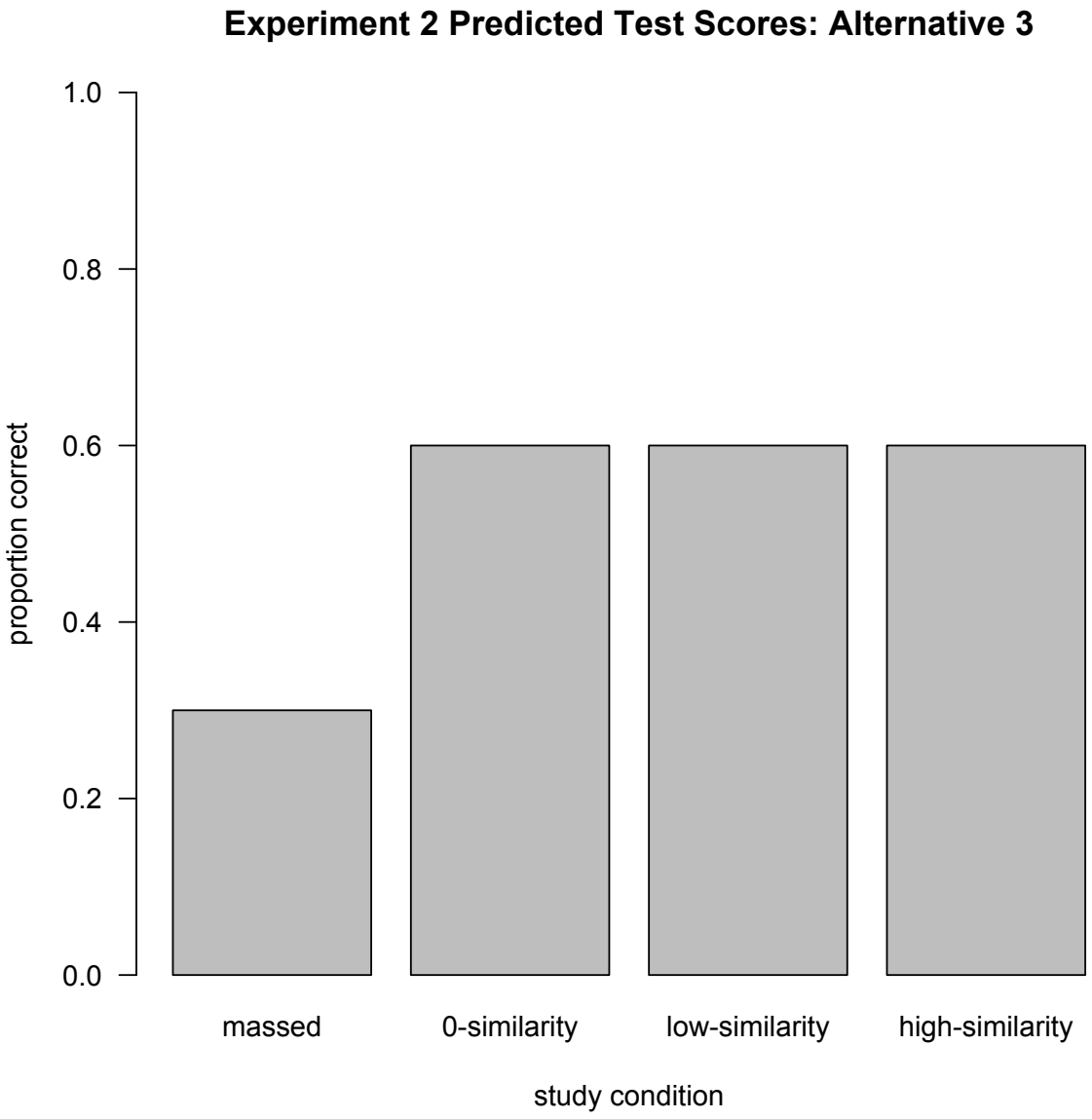


Figure 10. Bar graphs depicting predicted results for Experiment 2 based on the hypothesis that interleaving either similar or dissimilar materials should result in the same level of categorization test performance.

Results

Proportion correct. Figure 11 shows a summary of the categorization test results for the proportion of correct responses. The mean proportion of correct responses across all four conditions (M = .727, SD = .300) was relatively high compared with Experiment 1 (M = .345, SD = .123). This may have been due to the fact that participants viewed each stimulus in the study phase twice in Experiment 2 rather than once as in Experiment 1, and it may also be due to the smaller number of categories participants had to learn (eight versus twenty-four). A repeated measures ANOVA with study condition as a within-subjects factor was conducted to compare the effect of study condition on the proportion of correct responses in massed, interleaved-0 similarity, interleaved-low similarity, and interleaved-high similarity conditions. There was not a significant effect of study condition on the proportion of correct responses on the test at the $p < .05$ level, $F(3, 183) = 1.00$, $MS_e = .050$, $p = .393$.

Figure 11

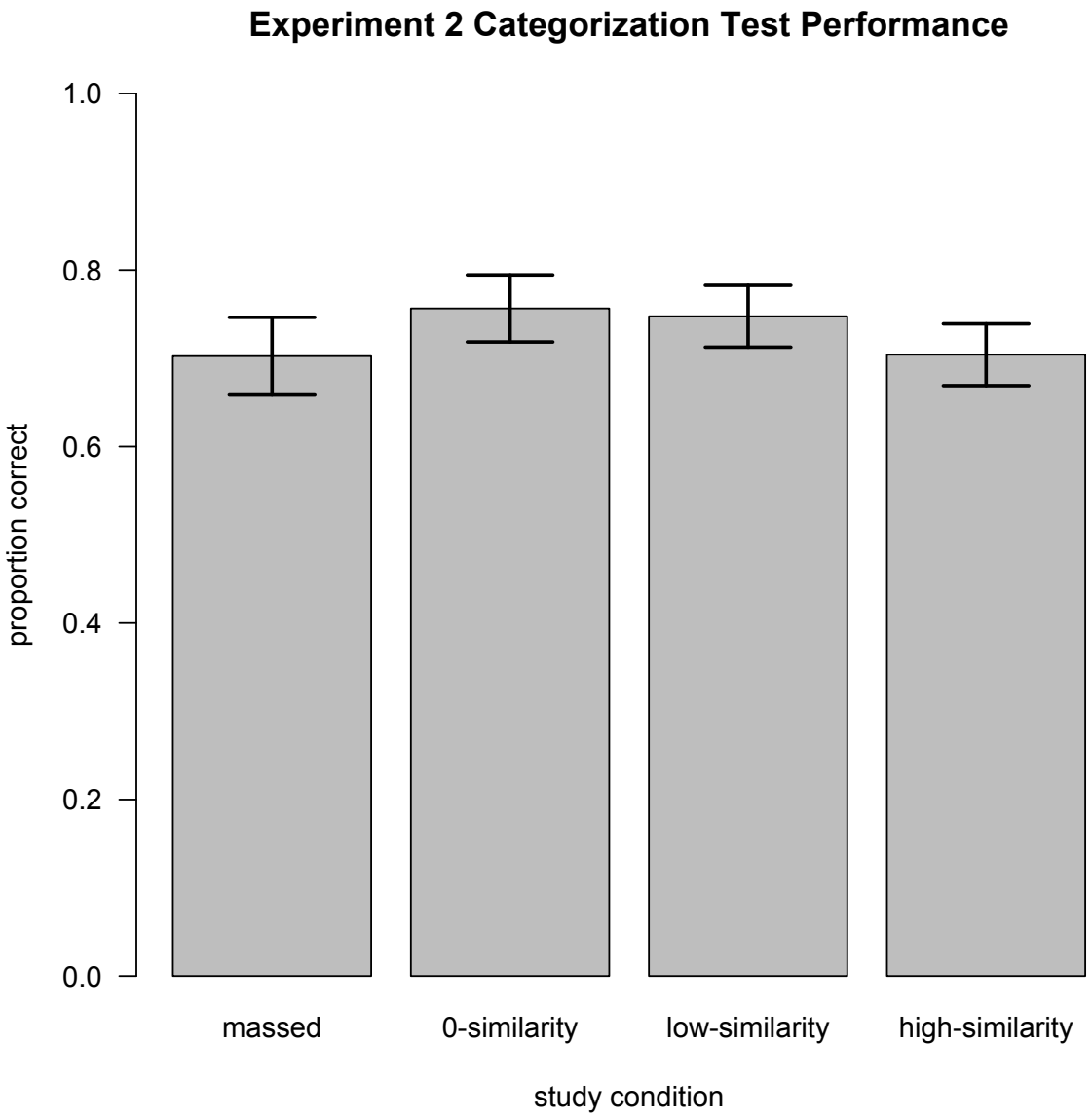


Figure 11. Bar graph depicting the mean proportion of correct responses on the 80-item categorization test for Experiment 2 (N = 62). Error bars represent the standard error of the mean. Differences between means were not significant at the $p < .05$ level.

Response times. Figure 12 shows a summary of the categorization test results for response times in milliseconds for trials with correct responses only. Response times that fell above 2.5 standard deviations of a participant's mean response time or were below 200 milliseconds were discarded from the analyses. This resulted in the omission of 3.2% of total trials. In addition, five participants were lacking any correct trials with valid response times in one-to-two of the four study conditions, and data for six of the 248 cells in the response time matrix were replaced with the mean response time for the cell's study condition in order to conduct the subsequent analyses. The overall mean response times for Experiment 2 (M = 2210, SD = 1087) were relatively fast compared with Experiment 1 (M = 5412, SD = 1671), which may partly reflect studying stimuli twice and/or the smaller number of response alternatives in Experiment 2. The speeding of response times is consistent with Hick's law, which states that response times increase logarithmically with the number of response alternatives (Hick, 1952). There was not a significant effect of study condition on test performance at the $p < .05$ level for the four study conditions, $F(3, 183) = .84$, $MS_e = 805{,}146$, $p = .471$.

Figure 12

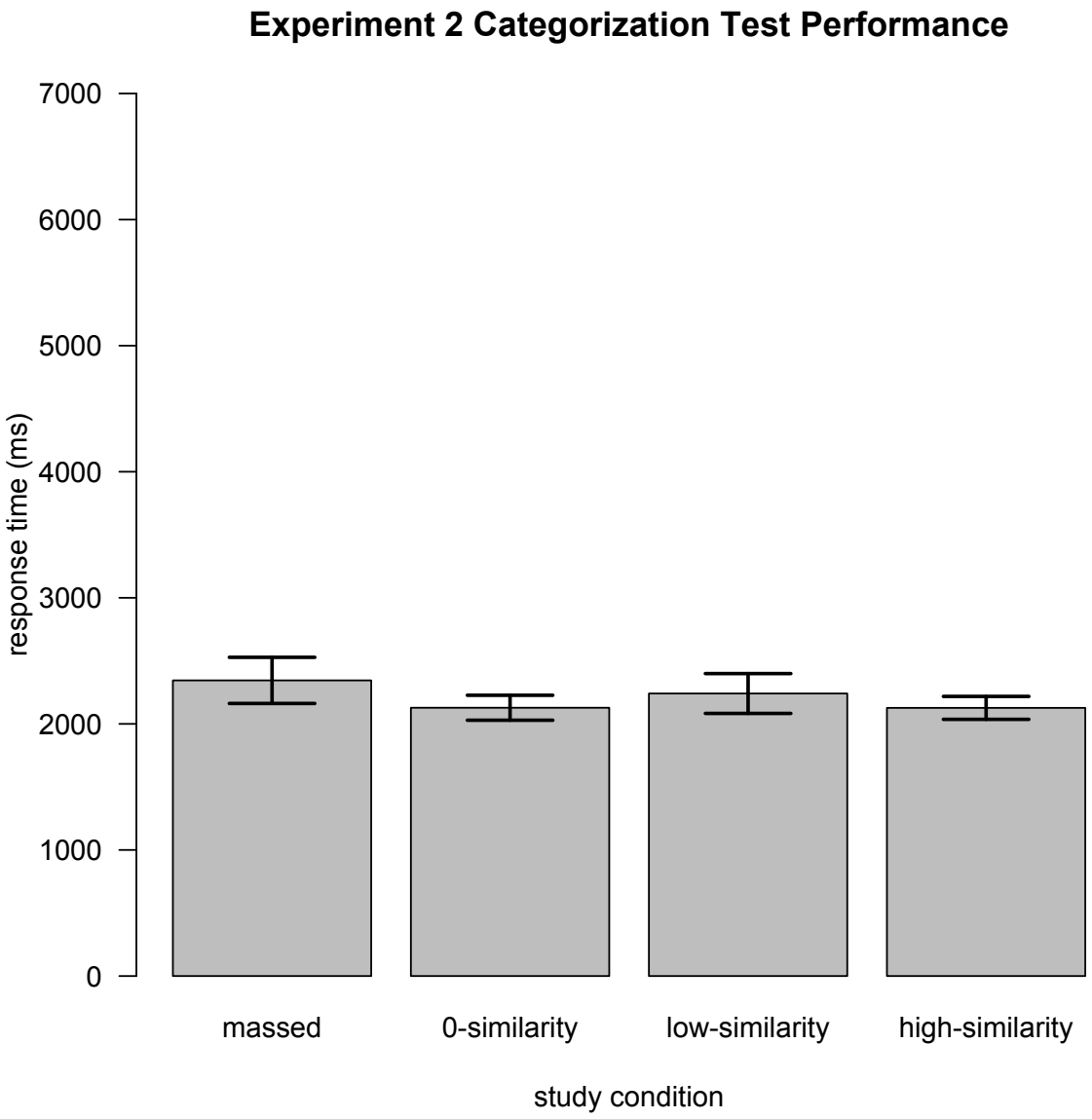


Figure 12. Bar graph depicting the mean response times for correct trials on the 80-item categorization test for Experiment 2 (N = 62). Error bars represent the standard error of the mean. Differences between means were not significant at the $p < .05$ level.

Post-experiment survey. After completing the experiment participants were given a two-item survey. The first item explained how some of the stimuli in the study phase of the experiment were presented in an interleaved versus massed way, and the participants were instructed to circle "yes" or "no" that they had noticed this difference. 83.9% of participants indicated that they noticed while the remaining 16.1% indicated that they did not notice. For the second item participants were instructed to circle which of the two study conditions they thought had helped them learn best, "massed" or "interleaved". Thirty-two percent of participants indicated that massed studying was more helpful, while the remaining sixty-eight percent indicated that interleaved studying was more helpful. The participants were not surveyed on their familiarity with the stimuli as it was assumed that the plaid images were novel to all the participants.

Discussion

In contrast to the results of Experiment 1, the results of Experiment 2 did not demonstrate significant interleaving effects in the scores on the categorization test, nor were there significant differences based on the similarity of interleaved materials. The mean response times on the categorization test were also not significantly different by study condition; however, as in Experiment 1 mean response times for this experiment should be interpreted with some caution due to the complexity of making a response on the keyboard to eight possible alternative answers.

Why was there such a difference in the pattern of results across Experiments 1 and 2? One possibility is that the way in which stimulus similarity was defined was quite different across the two experiments. In one sense the similarity was controlled most effectively in Experiment 2 by systematically varying the spatial frequencies of the visual

information at different orientations in the images. However, from a subjective point of view the there might be more similarity amongst each of the eight plaid categories than there is amongst the twelve categories of birds or paintings. That is, even two categories that were paired together as "interleaved-0 similarity" (e.g., plaid category LLH and plaid category HHL) in Experiment 2 might have shared more subjective visual similarity with each other than two categories that were paired together as "interleaved-dissimilar" (e.g., a Warbler and a Seurat painting) in Experiment 1. It may be argued, therefore, that the range of stimulus similarity sampled in Experiment 1 was larger than the range of stimulus similarity sampled in Experiment 2. This could have resulted in the lack of differences in test scores between the three levels of similarity in Experiment 2, although it does not address why no interleaving effects were found overall relative to the massed condition. How to best quantify similarity is an issue that should be addressed in follow-up experiments.

Another difference between Experiment 1 and Experiment 2 that could also have led to the different patterns of results was the difference in presentation time for each stimulus in the study phase. The presentation time for Experiment 2 (three seconds) was chosen to be the same as in Kornell & Bjork (2008), who used paintings as stimuli. However, the presentation time for Experiment 1 (six seconds) was increased in order to be closer in line with Wahlheim et al. (2011) who used birds as stimuli (the presentation time in that study was eight seconds). It is possible that participants did not have enough time to sufficiently view each stimulus, resulting in poor encoding processes. This may have led to the initially low categorization test performance during piloting that necessitated modifying the experiment procedure to repeat the study phase.

The repetition of the study phase presents an additional confound between Experiments 1 and 2. While presenting each stimulus twice in Experiment 2 accomplished the goal of raising performance on the categorization test above the floor, it also affected the manipulation of study condition. In effect, the massed condition was also a spaced condition, albeit one in which there were only two blocks of spaced study. As a result, the modified study design may have watered down the differences between the massed and interleaved conditions and made it more difficult to detect any interleaving effects that may have otherwise been observable.

Another possible explanation for the different patterns of results between Experiments 1 and 2 is that a much smaller number of categories were used in Experiment 2 than in Experiment 1, and this likely led to the observed greater categorization performance in Experiment 2 than in Experiment 1. The benefits of interleaving may be restricted to tasks that are relatively difficult. When tasks are easy, performance may be too close to the ceiling to find interleaving effects.

Overall, the results of Experiment 2 do not replicate the results of Experiment 1 or the results of many other studies demonstrating the benefit of interleaved over massed practice for long-term memory. The lack of differences observed between the three interleaved conditions in Experiment 2 also precludes an answer to the main question of the current study in terms of the hypotheses discussed, which is whether or not interleaving effects should be stronger or weaker depending on the similarity of the interleaved information. Several differences between the designs of Experiments 1 and 2 may explain the differing patterns of results and follow-up experiments using these

abstract visual stimuli should address aspects of the experimental design that may have contributed to the null results observed.

Experiment 3

One goal of Experiment 3 was to explore whether interleaving effects persist in classroom environments. The interleaving effect has been examined in very few published classroom studies to date (Rau et al., 2010; Rohrer et al., 2013), and even though the retention intervals in both these studies were longer than a day the results are conflicting. An important question remains, then, of whether interleaving effects generalize to classroom environments at all. Furthermore, the materials used in that study were graphical depictions of fractions problems. It may be possible that interleaving mathematics materials produces different learning patterns than interleaving other types of materials commonly used in classrooms.

The current study investigated interleaving effects using textual materials and college students as participants. A recently published study by Zulkiply, McLean, Burt, and Bath (2012) found significant interleaving effects at a brief retention interval when participants learned to categorize different psychopathologies by reading case study examples in the laboratory. This suggests that the benefit of interleaved practice extends beyond just categorizing visual images or solving math problems. Another goal of Experiment 3 was the same as Experiment 2, to investigate whether interleaving effects may be stronger or weaker depending on the similarity of the interleaved information, using stimuli that permit similarity to be controlled and systematically varied.

Participants

One-hundred and two participants (78 female, age: $M = 22.4$ $SD = 6.1$) from seven college courses at the University of Minnesota and Macalester College participated in the experiment either as part of their course curriculum or for extra credit. Participants from Macalester College were recruited from two introductory psychology courses, while participants from the University of Minnesota were recruited from five undergraduate- and graduate-level educational psychology courses.

Design and Materials

The design for Experiment 2 was very similar to Experiment 3, with the exception of the materials used. Participants in this experiment were instructed that their task was to learn to categorize a set of individuals based on personality type. They were given several worksheets during the study phase containing short profiles of hypothetical individuals with different personality types, and were given a new set of personality profiles to categorize at test. As in Experiment 2, eight unique categories were created corresponding to eight different personality types. Also similar to Experiment 2, the independent variable of *study condition* contained four levels (*massed, interleaved-0 similarity, interleaved-low similarity, and interleaved-high similarity*) and was manipulated within subjects.

However, an additional between-subjects variable of *retention interval* with two levels (*immediate* and *two days*) was introduced in Experiment 3 in order to test hypotheses about when interleaving effects emerge after study. Some prior studies such as Taylor and Rohrer (2010) showed that interleaving effects only emerge after a twenty-four hour delay between study and test, and that massed study produces superior test

performance at brief retention intervals. Because one of the goals of Experiment 3 was to explore the application of interleaving effects to classroom environments, a two-day retention interval condition was introduced. Thus, four of the seven classrooms that participated in the experiment were in the immediate retention interval condition and took the categorization test roughly one minute after completing the study phase, and three of the seven classrooms were in the two-day retention interval condition and took the categorization test two days after completing the study phase. The dependent variables were the proportion of novel stimuli correctly categorized at test and the mean response time for test items scored as correct.

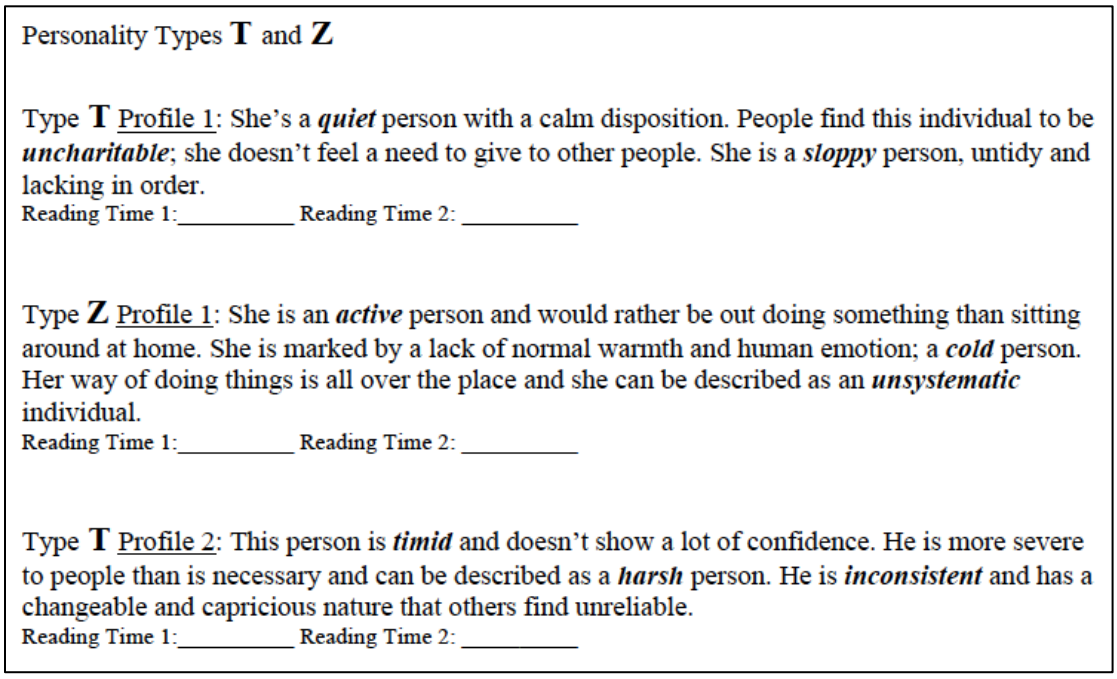
The study and test materials consisted of eighty three-sentence personality profiles, forty-eight of which were presented in the study phase and thirty-two in the test phase. Ten profiles were created for each of the eight categories. The three "big five" personality factors of extraversion, agreeableness, and conscientiousness were used to create the same three-feature similarity manipulation as Experiment 2, such that the eight personality types were defined as different combinations of positive or negative (high versus low) levels of the three different factors. With high or low levels of each of the three factors, the eight personality traits were : HHH, LLL, HHL, HLL, HLH, LHL, LLH, and LHH. For example, a personality trait with high extraversion, high agreeableness, and high conscientiousness was HHH.

Profiles for each individual were created using the personality trait adjectives developed by Goldberg (1992). In this set of adjectives, ten positively loading and ten negatively loading adjectives are associated with each big-five factor. One adjective from each of the three big five traits was selected to create each profile. For instance, one

profile for an HHH individual contained the words “daring, considerate, and systematic” while another profile for an HHH individual contained the words “talkative, cooperative, and efficient.” Each profile was comprised of three full sentences containing the three critical adjectives. The sentences were created using definitions of the adjectives taken from four online dictionaries (Wiktionary, Merriam-Webster, Dictionary.com, and the Cambridge Free English dictionary). As in Experiment 2, the letters S, T, U, V, W, X, Y, and Z were used to designate each of the eight categories. The category letter name was printed next to each profile in bold typeface during the study phase. See Figure 13 for an example of the profile worksheets used.

The definitions of similarity were identical to Experiment 2, except that the stimuli in Experiment 3 shared semantic rather than visual features. Thus, two 0-similarity personality types were defined by sharing opposite loadings on the three big five traits (e.g., high extraversion, high agreeableness, and low conscientiousness versus low extraversion, low agreeableness, and high conscientiousness). Two low-similarity personality types shared the same loading on one personality feature, while two high-similarity personality types shared the same loading on two personality features.

Figure 13


Personality Types **T** and **Z**

Type **T** Profile 1: She's a ***quiet*** person with a calm disposition. People find this individual to be ***uncharitable***; she doesn't feel a need to give to other people. She is a ***sloppy*** person, untidy and lacking in order.
Reading Time 1:_____ Reading Time 2: _____

Type **Z** Profile 1: She is an ***active*** person and would rather be out doing something than sitting around at home. She is marked by a lack of normal warmth and human emotion; a ***cold*** person. Her way of doing things is all over the place and she can be described as an ***unsystematic*** individual.
Reading Time 1:_____ Reading Time 2: _____

Type **T** Profile 2: This person is ***timid*** and doesn't show a lot of confidence. He is more severe to people than is necessary and can be described as a ***harsh*** person. He is ***inconsistent*** and has a changeable and capricious nature that others find unreliable.
Reading Time 1:_____ Reading Time 2: _____

Figure 13. An example of part of a worksheet for the *interleaved-high-similarity* study condition, in which profiles for personality types LLL and HLL were interleaved.

Six profiles for each personality type served as study stimuli and four served as test stimuli. The four study conditions were presented as individual worksheets, so that each participant read through eight worksheets, two per study condition. A single worksheet contained six profiles, which were ordered according to study condition. A massed worksheet contained six profiles from one personality type category. Each of the worksheets for the three interleaved study conditions contained three profiles from one category alternating with three profiles from another category. At the bottom of each worksheet was a seven-point confidence rating scale for the participant to evaluate his or her ability to correctly categorize a new profile based on what was just read on that study worksheet. This was done in order to encourage category induction and ensure sufficient encoding of the stimuli.

Participants in the immediate retention interval group completed a simple maze task on a worksheet before proceeding to the categorization test, while participants in the two-day retention interval group were not given a maze task and took the categorization test during the next class period. The categorization test was administered on paper and contained thirty-two new profiles, four each from the eight personality types presented in pseudo-random order. Under each profile was printed the eight category letters to which that personality profile might correspond.

Because the presentation time of each stimulus was not controlled as in Experiments 1 and 2, a rough measure of reading time during the study phase and response time during the test phase was also collected. An online stopwatch website was used to start a clock timer at the beginning of the experiment and was projected on a screen to the class. Under each profile on the study and test worksheets there was a space in which participants were instructed to print the time on the clock when they completed reading each profile during the study phase. These data were also used as a manipulation check to ensure that participants were reading each of the profiles in the order they were presented on the worksheets. Similar to Experiment 2, the participants went through all stimuli in the study phase twice, so there were two spaces provided under each profile where they were to indicate the clock time after reading it. Participants were also instructed to print the clock time in the space provided after they completed each item on the categorization test. These data were to be used as a coarse measure of response time in the analysis of categorization test performance.

Procedure

The experiment took place during one class period for classes in the immediate retention interval condition and over two class periods in the two-day retention interval condition. At the beginning of the study (day one for the two-day retention interval group), the experimenter passed out the instructions and asked the class to read them and focus their attention on the experimenter at the front of the room when they were finished. The experimenter then verbally summarized the instructions and gave the class an opportunity to ask questions about them. The experimental materials were then passed out to the class–the immediate retention interval classes received the study and test materials in one packet, while the two-day retention interval classes received just the study materials.

Participants were instructed to read through each profile on the worksheets one time only, in the order they were presented, and to complete the confidence rating at the bottom of each worksheet when it was reached. Similar to Experiment 2, the results of initial piloting on this procedure showed that categorization test performance was close to floor (not significantly above chance at .125), indicating that the task was rather difficult. Because of this, the procedure was modified so that participants were instructed to read through all eight worksheets, then go back and start from the beginning and read them through a second time to encourage sufficient encoding. They were also instructed that they could change their confidence ratings during the second reading of the worksheets if they thought their ratings had changed.

Students in the two-day retention interval condition read through the forty-eight study profiles twice in succession on day one and were permitted to leave the class when

they were finished. When they returned to class on day two, they were given a new set of instructions for the categorization task using the same procedure described above for day one. Then the test was passed out along with a post-experiment survey. Participants were instructed to read each profile on the test and circle the letter corresponding to the personality type described by that profile, as well as print the clock time after completing each item. After completing the test and the post-experiment survey, students were instructed to bring their completed study documents to the experimenter and then sit quietly until the rest of the class was finished and the class lesson for the day then began. However, for the group of students from one University of Minnesota course who participated in the two-day retention interval condition, the procedure was done outside of the regular classroom environment and so students were permitted to leave upon completing the experiment.

Students in the immediate retention interval also read through the forty-eight study profiles twice in succession, but then completed a brief maze worksheet before proceeding to the categorization test. After completing the test and the post-experiment survey students were permitted to leave class. The full procedure in both retention interval conditions took about an hour to complete.

Predictions

The predictions for Experiment 3 were identical to Experiment 2 with respect to study condition (see Figures 8-10). With respect to retention interval, it was hypothesized that any interleaving effect would be greater for the two-day retention interval condition than the immediate retention interval condition. Based on the results of Helsdingen et al. (2011) and Taylor and Rohrer (2010), it was further hypothesized that the massed

condition may produce better performance on the categorization test in the immediate retention interval condition.

Results

Proportion correct. Figures 14 and 15 show a summary of the categorization test results for the proportion of correct responses in each retention interval group. Four participants from the two-day retention interval condition did not return on day two. One participant's test data was excluded for procedural errors, and two participants' data were excluded because they were not naïve to the experiment protocol. Of the remaining ninety-five, fifty-one additional participants were excluded from the analyses because their mean test scores were not significantly above chance performance (*t*s < 1); thirteen of those participants were in the immediate retention interval group while the remaining thirty-seven were in the two-day retention interval group. The resulting *N* for the analysis of variance reported below was 44, with twenty-one participants in the immediate retention interval group and twenty-three participants in the two-day retention interval group. The mean proportion of correct responses across all conditions for the participants who performed above chance was .505 (*SD* = .314).

A 2 X 4 mixed ANOVA with study condition (massed, interleaved-0 similarity, interleaved-low similarity, interleaved-high similarity) as a within-subjects factor and retention interval (immediate, two-day) as a between-subjects factor was conducted to compare the effect of study condition and retention interval on the proportion of correct responses. The main effect of study condition on the proportion of correct responses did not approach significance at the $p < .05$ level, $F(3, 126) = .70$, $MS_e = .057$, $p = .552$. Neither the main effect of retention interval, $F(1, 42) = .76$, $MS_e = .233$, $p = .388$, nor the

interaction between study condition and retention interval, $F(3, 126) = .23$, $MS_e = .057$, $p = .874$, approached significance. Because such a considerable proportion of the participant sample was excluded, the analysis of variance was repeated including participants whose performance was significantly below chance. The main effect of study condition was again non-significant, $F(3, 276) = 1.40$, $MS_e = .038$, $p = .243$, as was the main effect of retention interval, $F(1, 92) = 1.56$, $MS_e = .060$, $p = .215$, and the interaction between study condition and retention interval, $F(3, 276) = .761$, $MS_e = .038$, $p = .517$.

Figure 14

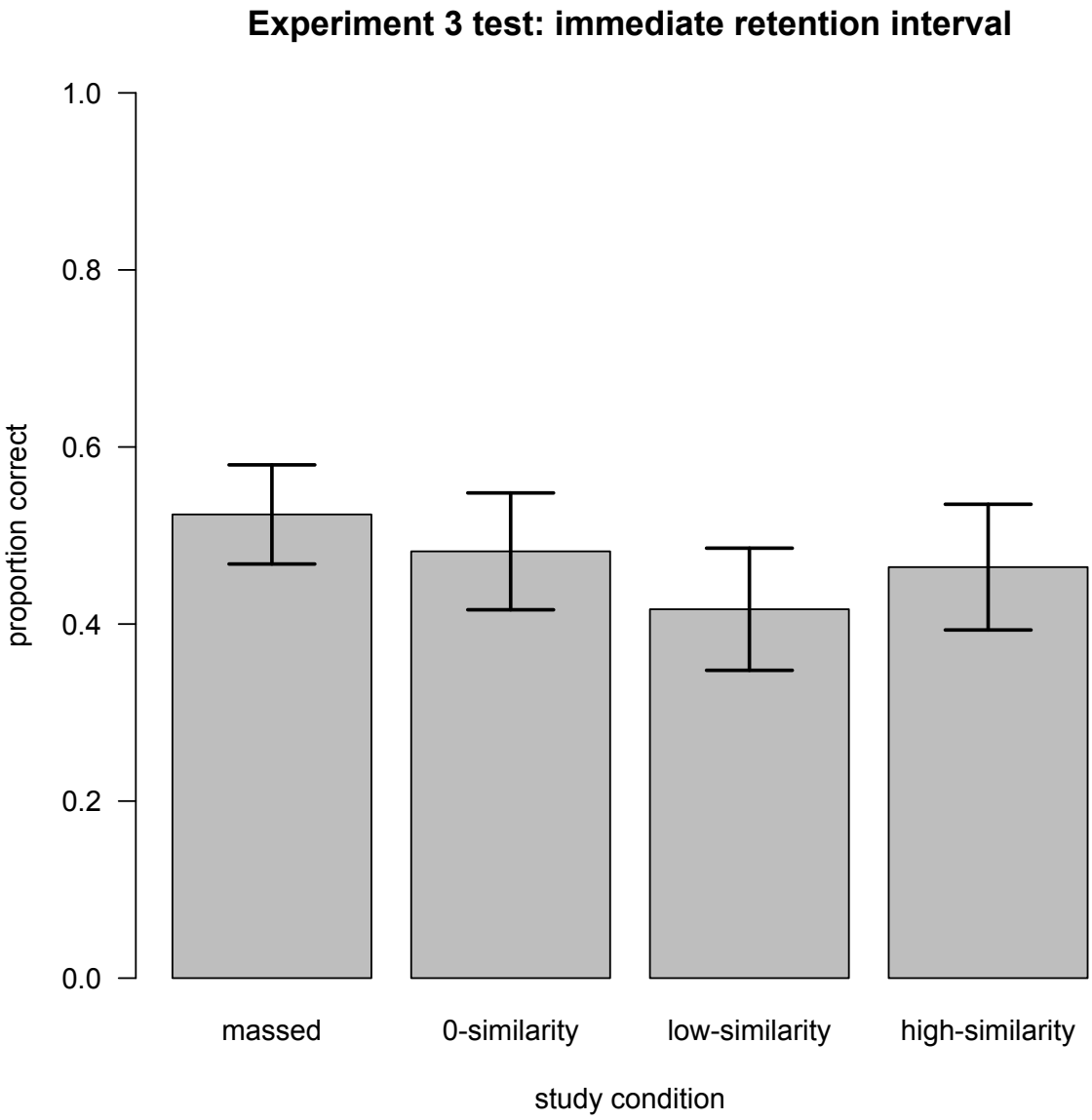


Figure 14. Bar graph depicting the mean proportion of correct responses on the 32-item categorization test for Experiment 3 for the immediate retention interval group (N = 21). Error bars represent the standard error of the mean. Differences between means were not significant at the $p < .05$ level.

Figure 15

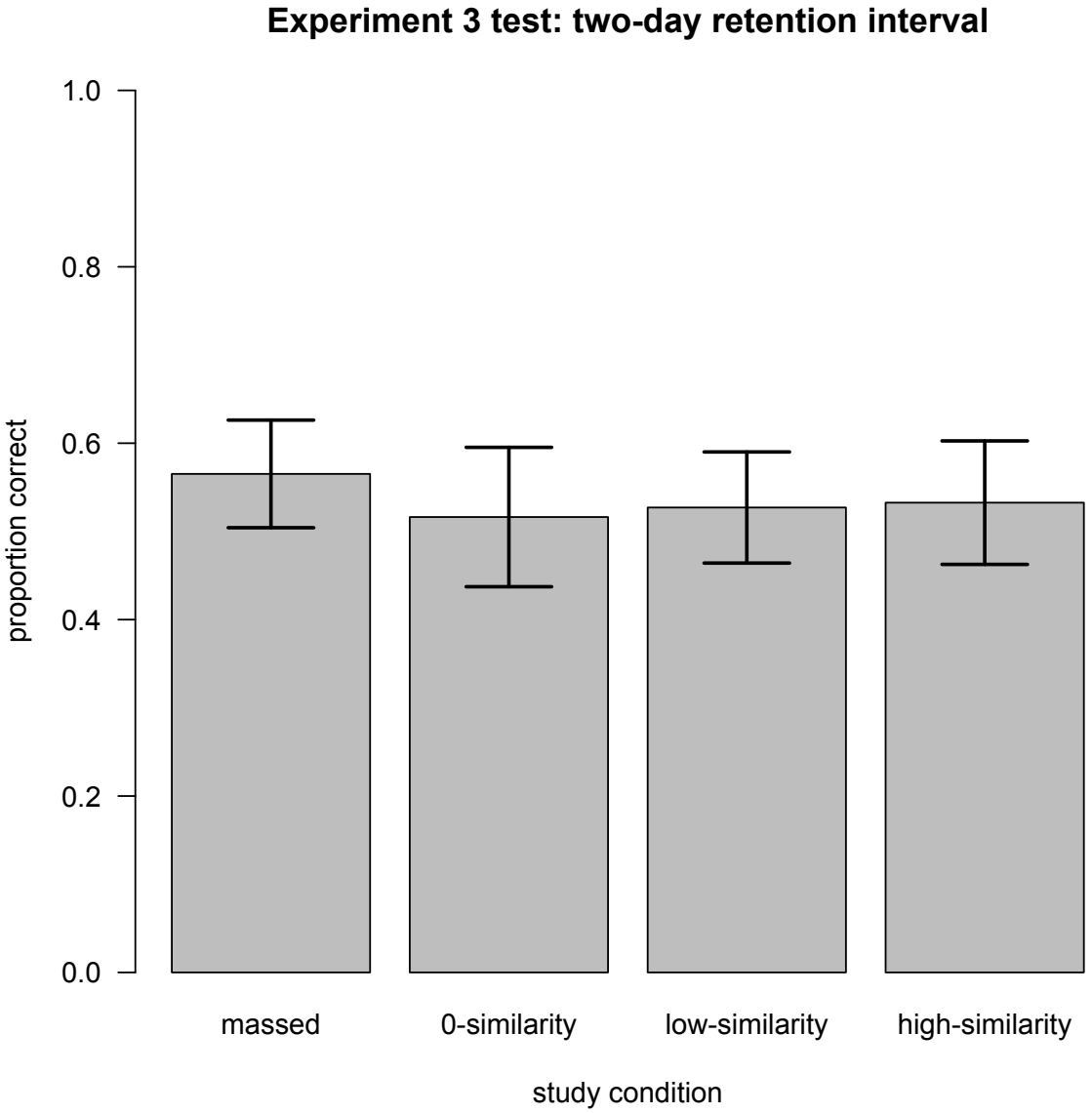


Figure 15. Bar graph depicting the mean proportion of correct responses on the 32-item categorization test for Experiment 3 for the two-day retention interval group (N = 23). Error bars represent the standard error of the mean. Differences between means were not significant at the $p < .05$ level.

Response times. Due to the relatively small number of trials per study condition on the categorization test (8) and the relatively low test scores (M = .33), there were too few correct trials with valid response times to make a formal analysis feasible. Therefore, response time data were not analyzed for Experiment 3.

Study times. Experiment 3 was unlike Experiments 1 and 2 in that the duration of the study phase was not fixed–participants read through the profiles at their own pace. The clock times recorded by the participants when reading the study profiles were used to compare the total time spent studying the profiles with their scores on the categorization test. The mean study time was 28:24 minutes (*SD* = 6:11). The comparison between mean study time for the immediate retention interval group (*M* = 27:57, *SD* = 6:36) and the two-day retention interval group (*M* = 28:38, *SD* = 5:59) did not reach significance ($t < 1$). Overall, there was a significant positive correlation between time spent studying the profiles and categorization test score, $r(93) = .29$, $p < .05$.

Post-experiment survey. After completing the experiment participants were given a three-item survey. The first item explained how some of the stimuli in the study phase of the experiment were presented in an interleaved versus massed way, and the participants were instructed to circle "yes" or "no" that they had noticed this difference. Ninety percent of participants indicated that they noticed while the remaining ten percent indicated that they did not notice. For the second item participants were instructed to circle which of the two study conditions they thought had helped them learn best, "massed" or "interleaved". Sixty-seven percent of participants indicated that massed studying was more helpful, while thirty-two percent indicated that interleaved studying was more helpful (one participant did not respond). The last item on the survey inquired

into participants' prior knowledge of the concept of the big five personality traits that were used in the experiment. They were instructed to indicate their level of familiarity with the big five on a three-point scale: "unfamiliar", "somewhat familiar", and "very familiar". Forty-three percent of participants indicated that they were unfamiliar with the big five, thirty-nine percent indicated that they were somewhat familiar, and eighteen percent indicated that they were very familiar. While the unequal numbers of participants who responded to the three different alternative answers to item three on the survey make pairwise comparisons problematic, mean test scores for participants who were very familiar with the big five (M = .545, SD = .340) were numerically higher than test scores for participants who were somewhat familiar (M = .296, SD = .232) or unfamiliar (M = .286, SD = .218) with the big five.

Discussion

The results of Experiment 3 failed to replicate interleaving effects found in prior published laboratory-based studies and no significant differences in categorization test performance were found by either study condition or retention interval. It may be useful to consider differences between the current experiment and other studies in which significant interleaving effects were found at both brief and longer retention intervals.

Perhaps the most notable result of Experiment 3 was the large proportion of participants who did not score above chance on the categorization test. The difficulty of learning to categorize the textual materials used was apparently greater compared with the difficulty of categorizing materials in other published studies, or compared with the difficulty of categorizing the materials in Experiments 1 and 2 of the current study. This increased level of difficulty introduced an additional variable into the results, which was

that 62% of participants in the two-day retention interval did not score significantly above chance on the categorization test, while only 38% of those in the immediate retention interval did not score above chance. It was expected that test scores for the two-day retention interval group would be lower overall than for the immediate retention interval group due to forgetting, but it is problematic that well over half of the data from the two-day retention interval group was discarded from the analyses.

One key feature of Experiment 3 that may have contributed to the overall difficulty of learning the categories was the complexity of the materials used. In order to accomplish the task it was important that participants ignore similarity based solely on the specific trait adjective used. For instance, one profile for the personality type HHH used in either the study or test phase contained the adjective “talkative”, which was one of the ten positively loading adjectives for extraversion from Goldberg (1992). However, one profile for personality type HLL also contained the word “talkative”, as did one of the profiles for personality types HHL and HLH. Participants needed to learn that the combination of trait loadings on each profile determined whether or not two profiles described the same personality type and to avoid being lured by the simple overlap between specific adjectives used. This feature of the experiment most likely contributed to the difficulty. Despite that, there were a small number of participants in the sample who got perfect scores on the categorization test, which means that it was at least possible to use the information in the profiles to effectively learn the features of each personality type.

Another potentially critical difference between Experiment 3 and the other experiments in the current study is the use of textual versus visual materials. Interleaving

effects may depend in part on the nature of the information to be studied. Identifying visual features that are diagnostic to a particular category of images may be quite a different process cognitively than identifying semantic features in language that are diagnostic to a semantic category. Learning visual categories may be easier than learning textual categories, given our relative expertise with visual processing compared with reading. The visual features that make each category of images unique may be processed in a more automatic way while understanding the semantic features that make each personality type unique may require more deliberate processing. However, given that significant interleaving effects were found using textual descriptions of crimes in Helsdingen et al. (2011) and textual case studies in Zulkiply et al. (2012), the difficulty may have been due to the actual content of the materials created for Experiment 3 rather than the fact that they were presented in a textual medium.

A drawback to the design of Experiment 3 was the lack of random assignment to the retention interval conditions, which was due to the logistics of running the study in classrooms. Because of the need to work with specific course schedules, both courses that the author taught at Macalester college were assigned to the two-day retention interval group. The experimenter was thus the instructor of the course and the study was introduced as part of the course curriculum on memory or personality. The classes that were assigned to the immediate retention interval were all from the University of Minnesota and taught by instructors other than the author, which means that both school and instructor were confounded with retention interval in this experiment (there were also a small number ($N = 7$) of participants from the original sample in the two-day retention interval group from a University of Minnesota course). It is possible that differences in

the student populations between the two schools exist. It is also possible that participating in the experiment when it was part of the course curriculum and administered by the instructor could have impacted those participants' motivation in following the protocol and persisting at the task. Because the test scores do not differ significantly between the two retention interval groups its not possible to make any claims with regard to differences in motivation based solely on the results of this experiment.

The last aspect of Experiment 3 that may have affected the results is the relatively uncontrolled nature of the experiment protocol relative to Experiments 1 and 2. This is a natural consequence of conducting research in the classroom rather than the laboratory. While the order of stimulus presentation was highly controlled in the laboratory experiments in the current study, participants in Experiment 3 were able to peruse the study materials as they wished. While strict instructions were given to read each profile only once and in the order it was presented on the worksheet, the clock times recorded by the participants indicate that some of them were not complying with this instruction and were reading the profiles out of order. Because interleaving effects depend on the exact temporal sequencing of the stimuli, any failure to adhere to the order that the materials in the experiment were arranged in would have impacted the results negatively.

While it is inevitable that less experimental control is possible in a classroom-based study than a laboratory-based study, some modifications to the instructions in future follow-up experiments may help ensure that all participants are following the procedure more closely. Confusion about what information in the profiles is important for identifying each category may also be ameliorated through changes to the instructions.

As previously mentioned, because the specific adjectives used in the profiles were repeated across multiple personality types/categories, participants may have assumed that each adjective was diagnostic for a particular personality type. Instructions could make it clear that the combination of the three adjectives defines each personality type and that participants should ignore similarity based on the overlap between the specific adjectives used in the profiles, as well as emphasize the importance of proceeding through the experiment materials in order.

Overall, Experiment 3 was a useful exploration of how to translate laboratory research on spacing and interleaving into a classroom environment. While the results were disappointing, there are opportunities to modify the design of Experiment 3 for future research in order to maximize the likelihood of detecting interleaving effects in the classroom.

General Discussion

The results of the three experiments in the current study do not point to a clear answer to the main question of the study, which was: *is the interleaving effect greater in magnitude when the interleaved information is similar or when the interleaved information is dissimilar?* Considering each of the experiments individually, Experiment 1 showed results that most closely align with prior published studies demonstrating interleaving effects within a single experimental session. The results of Experiment 1 showed that interleaving images of birds and paintings resulted in significantly better categorization test performance than massing them. Furthermore, interleaving highly similar information resulted in significantly higher test scores compared with interleaving dissimilar information. The superior test scores for the interleaved-similar condition

relative to the interleaved-dissimilar condition support the discriminative contrast hypothesis of interleaving, which proposes that interleaving benefits learning best when the interleaved information comes from low-contrast categories where there is a high degree of overlap in information between categories.

The results of Experiment 1 were not consistent with the hypotheses that interleaving dissimilar information should result in larger interleaving effects than interleaving similar information or that equal benefits should be derived from interleaving similar or dissimilar information. Explanations based on the attenuation of attention or retrieval practice predict that interleaving dissimilar information should be preferable to interleaving similar information, and transfer appropriate processing predicts that interleaving similar or dissimilar information should yield equal benefits. However, the significant advantage found in Experiment 1 for the interleaved-similar study condition lends support to the idea that spacing and interleaving effects may be driven by enhanced discrimination processes when studying categories, rather than enhanced attention to the stimuli, facilitated retrieval processes, or a closer match between study and test conditions.

The results of Experiment 2 and 3 did not replicate the results of Experiment 1 or prior published studies demonstrating interleaving effects. There were several differences in the designs of the three experiments that may have led to the different patterns of results observed. Several of these differences may be addressed through simple modifications to the designs for future follow-up experiments. One key difference of potential importance between Experiment 1 and Experiments 2 and 3 not previously discussed was the differences in the proportion of massed versus interleaved stimuli in

the study phase. Experiment 1 was a between-subjects design, and all of the stimuli studied by a single participant were from either the massed condition or one of the interleaved conditions. However, Experiments 2 and 3 were within-subjects designs, and so one-quarter of the stimuli studied by a single participant were from each of the four study conditions. Three out of four of these conditions were interleaved, so only one-fourth of the stimuli were presented in a massed manner. It's possible that the lower proportion of massed stimuli in Experiment 2 and Experiment 3 created a "pop-out" effect and made those stimuli easier to learn, resulting in the lack of interleaving effects observed. Follow-up experiments could utilize between-subjects designs for Experiments 2 and 3 to determine if the same pattern of results would be found when all of the stimuli are presented in the same interleaved or massed condition for each participant.

Another notable difference between the three experiments discussed previously was that the study phases in Experiments 2 and 3 were repeated in order to improve performance on the categorization test. However, this meant that stimuli in the massed condition were actually studied by participants in a spaced manner, because they saw two blocks of each category presented instead of one. This means that the overall lack of interleaving effects found in Experiments 2 and 3 could be a result of unintentionally enhanced learning in the massed condition, which was actually a version of spaced practice in which there were two blocks of spaced study several minutes apart. This would suggest that the designs for Experiments 2 and 3 should be modified so that each stimulus is only presented once in the study phase. Doing so would require additional modifications to the materials or procedure to encourage deeper processing of the stimuli and thus better encoding. These modifications could include increasing the presentation

time of each stimulus in Experiment 2 or asking participants to read each personality profile twice in a row in Experiment 3. Because categorization test performance was so low in Experiment 3 despite the repetition of the study phase, further modifications to the materials may be needed. One possibility is to remove the sentence verbiage from the personality profiles so that each profile consists solely of the three adjectives that define positive or negative loading on each of the big five traits. Simplifying the study materials in this way may focus participants' attention on the critical information in each profile and reduce interference from the other, potentially distracting information in the sentences.

The different number of categories to be learned was another difference between Experiment 1 and Experiments 2 and 3. There were twenty-four categories in Experiment 1 and only eight in Experiments 2 and 3. The smaller number of categories used in Experiments 2 and 3 may have led to the observed greater categorization performance in Experiments 2 and 3 (that is, when participants performed above chance in Experiment 3) than in Experiment 1. Little to no benefit of interleaving may occur when performance is somewhat high or the task is somewhat easy, and the benefit may only occur when performance is somewhat low or the task is somewhat difficult. The number of categories to be learned, the number of exemplars used within each category, and the nature of the materials used in studies of interleaving are all variables which may collectively determine the overall difficulty level of the experimental task. It could be argued that when this difficulty level is too low (e.g., in Experiment 2) or too high (e.g., in Experiment 3, when participants who did not perform above chance are included)

interleaving effects may not be found. Keeping the difficulty level in a "sweet spot" for learning is a concern to be addressed in the piloting of future experiment designs.

Another question of potential interest for future research is whether the interleaving effects found in Experiment 1 would persist at different retention intervals. Published studies (e.g., Kornell & Bjork, 2008; Wahlheim et al., 2011) using these stimuli have reported interleaving effects at brief retention intervals, but spacing and interleaving effects have been observed in some studies only at retention intervals longer than one day (e.g., Bloom & Shuell, 1981; Rohrer & Taylor, 2007; Taylor & Rohrer, 2010). Would the interleaving advantage observed in Experiment 1 remain if participants were tested one day after study? With the goal in mind of applying the findings from spacing and interleaving studies to educational settings, utilizing the study design from Experiment 1 with a longer retention interval may provide more useful evidence.

In conclusion, the results of three experiments exploring the impact of stimulus similarity on interleaving effects are mixed. Differing patterns of results across the laboratory-based Experiments 1 and 2 suggest that spacing and interleaving effects may depend at least in part on the difficulty of the materials to be studied and the experimental design. The null results of classroom-based Experiment 3 highlight the complexity of translating laboratory research to less controlled classroom environments. Future research should explore interleaved practice with varied materials and at varied retention intervals to elucidate the exact nature of the effects.

References

Ambridge, B., Theakston, A. L., Lieven, E. V., & Tomasello, M. (2006). The distributed learning effect for children's acquisition of an abstract syntactic construction. *Cognitive Development*, 21(2), 174-193.

Appleton-Knapp, S., Bjork, R. A., & Wickens, T. D. (2005). Examining the spacing effect in advertising: Encoding variability, retrieval processes and their interaction. *Journal of Consumer Research,* 32, 266–276.

Bahrick, H. P. (1979). Maintenance of knowledge: Questions about memory we forgot to ask. *Journal of Experimental Psychology: General,* 108, 296–308.

Bahrick, H.P., Bahrick, L.E., Bahrick, A.S., & Bahrick, P.E. (1993). Maintenance of foreign-language vocabulary and the spacing effect. *Psychological Science,* 4, 316–321.

Bahrick, H. P., & Hall, L. K. (2005). The importance of retrieval failures to long-term retention: A metacognitive explanation of the spacing effect. *Journal of Memory and Language*, 52(4), 566-577.

Balota, D. A, Duchek, J. M., & Logan, J. M. (2007). Is expanded retrieval practice a superior form of spaced retrieval? A critical review of the extant literature. In J. S. Nairne (Ed.), *The foundations of remembering: Essays in honor of Henry L. Roediger III* (pp. 83-106). New York: Psychology Press.

Balota, D., Duchek, J., Sergent-Marshall, S., Roediger, H. (2006). Does expanded retrieval produce benefits over equal-interval spacing? Explorations of spacing effects in healthy aging and early stage alzheimer's disease. *Psychology and Aging*, 21(1), 19–31.

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.

Bloom, K. C., & Shuell, T. J. (1981). Effects of massed and distributed practice on the learning and retention of second-language vocabulary. *The Journal of Educational Research*, 74(4), 245-248.

Bransford, J. D., & Schwartz, D. L. (1999). Rethinking transfer: A simple proposal with multiple implications. *Review of research in education*, *24*, 61-100.

Bruce, D., & Bahrick, H. P. (1992). Perceptions of past research. *American Psychologist,* 47, 319–328.

Carson, L. M., & Wiegand, R. L. (1979). Motor schema formation and retention in young children: A test of Schmidt's schema theory. *Journal of Motor Behavior,* 11(4), 247-251.

Carvalho, P. F., & Goldstone, R. L. (2012). Category structure modulates interleaving and blocking advantage in inductive category acquisition. In N. Miyake, D. Peebles, and R. P. Cooper (Eds.), *Proceedings of the 34th Annual Conference of the Cognitive Science Society* (pp. 186-191). Austin, TX: Cognitive Science Society.

Catalano, J. F., & Kleiner, B.M. (1984). Distant transfer in coincident timing as a function of variability of practice. *Perceptual and Motor Skills*, 58(3), 851-856.

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.

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.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Hillsdale, NJ: Erlbaum.

Cook, T. W. (1934). Massed and distributed practice in puzzle solving. *Psychological Review,* 41, 330–355.

Cull, W.L. (2000). Untangling the benefits of multiple study opportunities and repeated testing for cued recall. *Applied Cognitive Psychology*, 14, 215–235.

Cull, W. L., Shaughnessy, J. L., & Zechmeister, E. B. (1996). Expanding understanding of the expanding-pattern-of-retrieval mnemonic: Toward confidence in applicability. *Journal of Experimental Psychology: Applied*, *2*, 365-378.

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.

Dempster, F. N. (1989). Spacing effects and their implications for theory and practice. *Educational Psychology Review*, 1(4), 309-330.

Donovan, J. J., & Radosevich, D. J. (1999). A meta-analytic review of the distribution of practice effect: Now you see it, now you don't. *Journal of Applied Psychology*, 84(5), 795.

Dunlosky, J., Rawson, K. A., Marsh, E. J., Nathan, M. J., & Willingham, D. T. (2013). Improving students' learning with effective learning techniques promising directions from cognitive and educational psychology. *Psychological Science in the Public Interest*, 14(1), 4-58.

Ebbinghaus, H. (1913). *Memory* (H. A. Ruger & C. E. Bussenius, Trans.). New York: Teachers College. (Original work published 1885; paperback ed., New York: Dover, 1964).

Finley, J. R., Benjamin, A. S., Hays, M. J., Bjork, R. A., & Kornell, N. (2011). Benefits of accumulating versus diminishing cues in recall. *Journal of Memory and Language*, 64(4), 289-298.

Gagne, 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.

Goldberg, L. R. (1992). The development of markers for the Big-Five factor structure. *Psychological Assessment*, *4*(1), 26-42.

Grote, M. G. (1995). Distributed versus massed practice in high school physics. *School Science and Mathematics*, 95(2), 97-101.

Hall, K.G., Domingues, D.A., & Cavazos, R. (1994). Contextual interference effects with skilled baseball players. *Perceptual and Motor Skills*, 78(3), 835-841.

Helsdingen, A. S., Van Gog, T., & van Merriënboer, J. J. (2011). The effects of practice schedule on learning a complex judgment task. *Learning and Instruction*, 21(1), 126-136.

Hick, W. E. (1952). On the rate of gain of information. *Quarterly Journal of Experimental Psychology*, 4(1), 11-26.

Janiszewski, C., Noel, H., & Sawyer, A. G. (2003). A meta-analysis of the spacing effect in verbal learning: Implications for research on advertising repetition and consumer memory. *Journal of Consumer Research*, 30(1), 138-149.

Kahana, M. J., & Howard, M. W. (2005). Spacing and lag effects in free recall of pure lists. *Psychonomic Bulletin & Review*, 12(1), 159-164.

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

Keller, G.J., Li, Y., Weiss, L.W., & Relyea, G.E. (2006). Contextual interference effect on acquisition and retention of pistol-shooting skills. *Perceptual and Motor Skills*, 103(1), 241-252.

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., Eich, T., & Bjork, R. (2010). Spacing as the friend of both memory and induction in older adults. *Psychology and Aging*, 25(2), 498–503.

Kurtz, K. H., & Hovland, C. I. (1956). Concept learning with differing sequences of instances. *Journal of Experimental Psychology*, 51(4), 239.

Landauer, T. K., & Bjork, R. A. (1978). Optimum rehearsal patterns and name learning. In M. Gruneberg, P. E. Morris, & R. N. Sykes (Eds.), *Practical aspects of memory* (pp. 625-632). London: Academic Press.

Landin, D. K., Hebert, E. P., & Fairweather, M. (1993). The effects of variable practice on the performance of a basketball skill. *Research Quarterly for Exercise and Sport,* 64(2), 232.

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

Lee, T. D., & Genovese, E. D. (1988). Distribution of practice in motor skill acquisition: Learning and performance effects reconsidered. *Research Quarterly for Exercise and Sport*, 59(4), 277-287.

Lee, T. D., Magill, R. A., & Weeks, D. J. (1985). Influence of practice schedule on testing schema theory predictions in adults. *Journal of Motor Behavior*, 17(3), 283-299.

Morris, C. D., Bransford, J. D., & Franks, J. J. (1977). Levels of processing versus transfer appropriate processing. *Journal of Verbal Learning and Verbal Behavior*, 16(5), 519-533.

Morris, P. E., & Fritz, C. O. (2000). The name game: Using retrieval practice to improve the learning of names. *Journal of Experimental Psychology: Applied*, *6*, 124-129.

Moxley, S. E. (1979). Schema: The variability of practice hypothesis. *Journal of Motor Behavior,* 11(1), 65-70.

Newell, K. M., & Shapiro, D. C. (1976). Variability of practice and transfer of training: Some evidence toward a schema view of motor learning. *Journal of Motor Behavior*, 8(3), 233-243.

Pyc, M. A., & Rawson, K. A. (2007). Examining the efficiency of schedules of distributed retrieval practice. *Memory & Cognition*, 35(8), 1917-1927.

Rau, M. A., Aleven, V., & Rummel, N. (2010). Blocked versus interleaved practice with multiple representations in an intelligent tutoring system for fractions. In V. Aleven, J. Kay, & J. Mostow (Eds.) *Proceedings of the 10th International Conference of Intelligent Tutoring Systems* (pp. 413-422). Heidelberg / Berlin:

Springer.

Rea, C. P. & Modigliani, V. (1985). The effect of expanded versus massed practice on the retention of multiplication facts and spellings lists. *Human Learning: Journal of Practical Research and Applications*, 4, 11-18.

Rea, C. P., & Modigliani, V. (1987). The spacing effect in 4-to 9-year-old children. *Memory & Cognition*, 15(5), 436-443.

Reicher, G.M., Snyder, C.R.R., & Richards, J. T. (1976). Familiarity of background characters in visual scanning. *Journal of Experimental Psychology: Human Perception and Performance*, 1, 522-530.

Richland, L. E., Bjork, R. A., Finley, J. R., & Linn, M. C. (2005). Linking cognitive science to education: Generation and interleaving effects. In *Proceedings of the 27th Annual Conference of the Cognitive Science Society* (pp. 1850-1855). Mahwah, NJ: Erlbaum.

Roediger, H. L. (1990). Implicit memory: Retention without remembering. *American Psychologist*, 45(9), 1043-1056.

Rohrer, D. (2009). The effects of spacing and mixing practice problems. *Journal for Research in Mathematics Education*, 40(1), 4-17.

Rohrer, D. (2012). Interleaving helps students distinguish among similar concepts. *Educational Psychology Review,* 24(3), 355-367.

Rohrer, D., Dedrick, R. F., & Burgess, K. (2014). The benefit of interleaved mathematics practice is not limited to superficially similar kinds of problems. *Psychonomic Bulletin & Review*, 1-8.

Rohrer, D., & Pashler, H. (2007). Increasing retention without increasing study time. *Current Directions in Psychological Science*, 16(4), 183-186.

Rohrer, D. & Pashler, H. (2010). Recent research on human learning challenges conventional instructional strategies. *Educational Researcher,* 39(5), 406-412.

Rohrer, D. & Taylor, K. (2006). The effects of overlearning and distributed practise on the retention of mathematics knowledge. *Applied Cognitive Psychology*, 20(9), 1209-1224.

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

Seabrook, R., Brown, G. D., & Solity, J. E. (2005). Distributed and massed practice: From laboratory to classroom. *Applied Cognitive Psychology*, 19(1), 107-122.

Schmidt, R. A., & Bjork, R. A. (1992). New conceptualizations of practice: Common principles in three paradigms suggest new concepts for training. *Psychological Science*, 3(4), 207–217.

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(2), 179.

Snyder, K. A., Blank, M. P., & Marsolek, C. J. (2008). What form of memory underlies novelty preferences? *Psychonomic Bulletin & Review, 15*, 315-321.

Solity, J. E. (2000). The early reading research: Applying psychology to classroom practice. *Educational and Child Psychology,* 17(2), 46-65.

Tauber, S. K., Dunlosky, J., Rawson, K. A., Wahlheim, C. N., & Jacoby, L. L. (2013). Self-regulated learning of a natural category: Do people interleave or block exemplars during study? *Psychonomic Bulletin & Review*, 20(2), 356-363.

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

Wahlheim, C. N., Dunlosky, J., & Jacoby, L. L. (2011). Spacing enhances the learning of natural concepts: An investigation of mechanisms, metacognition, and aging. *Memory & Cognition*, 39(5), 750-763.

Wrisberg, C. A. & Ragsdale, M. R. (1979). Cognitive demand and practice level: Factors in the mental rehearsal of motor skills. *Journal of Human Movement Studies*, 5, 201-208.

Zulkiply, N. & Burt, J. S. (2013). The exemplar interleaving effect in inductive learning: Moderation by the difficulty of category discriminations. *Memory & Cognition*, 41(1), 16-27.

Zulkiply, N., McLean, J., Burt, J., & Bath, D. (2012). Spacing and induction: Application to exemplars presented as auditory and visual text. *Learning and Instruction*, 22(3), 215–221.