Effectiveness of Learner-Regulated Study Sequence: An *in-vivo* study in Introductory Psychology courses

Paulo F. Carvalho (pcarvalh@indiana.edu)
Department of Psychological and Brain Sciences, Indiana University, 1101 E. 10th Street Bloomington, IN 47405 USA

David W. Braithwaite (baixiwei@cmu.edu)
Department of Psychology, Carnegie Mellon University, 5000 Forbes Avenue, Pittsburgh, PA 15213 USA

Joshua R. de Leeuw (jodeleeu@indiana.edu)
Department of Psychological and Brain Sciences, Indiana University, 1101 E. 10th Street Bloomington, IN 47405 USA

Benjamin A. Motz (bmotz@indiana.edu)
Department of Psychological and Brain Sciences, Indiana University, 1101 E. 10th Street Bloomington, IN 47405 USA

Robert L. Goldstone (rgoldsto@indiana.edu)
Department of Psychological and Brain Sciences, Indiana University, 1101 E. 10th Street Bloomington, IN 47405 USA

Abstract

Study sequence can have a profound impact on learning. Previous research has often shown advantages for interleaved over blocked study, though the reverse has also been found. Learners typically prefer blocking even in situations for which interleaving is superior. The present study investigated learner regulation of study sequence, and its effects on learning in an ecologically valid context – university students using an online tutorial relevant to an exam that counted toward their course grades. The majority of participants blocked study by problem category, and this tendency was positively associated with subsequent exam performance. The results suggest that preference for blocked study may be adaptive under some circumstances, and highlight the importance of identifying task environments under which different study sequences are most effective.

Keywords: study sequence; *in-vivo* educational research; concept learning

Introduction

Learning is increasingly often taking place not in a passive context but in unsupervised situations in which learners must make active decisions about their own study. On the one hand, the increased opportunity for self-regulated study might include improved engagement and lead to better allocation of study time (Gureckis & Markant, 2012). On the other hand self-regulated learning might not lead to efficiency gains because of deficiencies of learners' knowledge about the efficacy of different study methods (Bjork, Dunlosky, & Kornell, 2013).

One important decision learners must make is how to sequence their study of different materials. Study sequence can have a profound effect on learning, even when the materials studied are kept constant (Elio & Anderson, 1984; Medin & Bettger, 1994). If these materials involve several examples from different concepts, learners might opt to block their study by studying all examples of one concept before studying a different concept. Alternatively, learners might choose to interleave examples from different concepts.

Numerous studies have found advantages for interleaved over blocked study sequences (Birnbaum, Kornell, Bjork, & Bjork, 2012; Kornell & Bjork, 2008; Rohrer & Taylor, 2007; Taylor & Rohrer, 2010; Wahlheim, Dunlosky, & Jacoby, 2011). For example, Kornell and Bjork (2008) showed that interleaved presentation of paintings from different artists resulted in better learning of the artists' styles and improved transfer to novel paintings, compared to presenting different artists' paintings in separate blocks. The opportunity afforded for comparison between successive examples may be critical to the effectiveness of interleaving (Birnbaum et al., 2012; Carvalho & Goldstone, 2014; Goldstone, 1996; Kang & Pashler, 2012). Studying successive examples from different categories maximizes between-category comparisons, which should facilitate learning of discriminative features.

However, sometimes the challenge is not so much finding differences between categories but finding commonalities between the different examples within each category. In these situations, blocked study, which maximizes within-category comparisons and thus highlights within-category similarities, may be more effective. Consistent with this view, several studies have found advantages for blocked study when between-category differences are obvious or within-category similarities subtle (Carpenter & Muller, 2013; Carvalho & Goldstone, 2014; Goldstone, 1996; Kurtz & Hovland, 1956; Zulkiply & Burt, 2012). For example, Goldstone (1996) demonstrated better learning of complex line pattern categories when they were studied in blocked rather than interleaved sequences. Critically, within-category similarities were subtle because different examples from a category only shared a few lines in common.

Yet, most studies looking at whether interleaved or blocked study sequences are more beneficial for learning use situations in which the learner does not choose how to sequence their study, while in everyday educational settings the student is often in control of how to organize their study. Indeed, Tauber et al. (2013) recently found that when learners could choose how to sequence examples of different categories during study, they overwhelmingly preferred to block the examples of each category. This is an interesting finding given that interleaved study has been shown to be more effective for learning the same type of stimuli (e.g., Wahlheim et al., 2011).

These results are consistent with previous evidence showing that when asked which learning sequence (interleaved or blocked) learners believe would result in best learning, the majority chooses blocked study (Kornell & Bjork, 2008; Zulkiply & Burt, 2012). One possible reason for this belief is that blocked sequences may facilitate processing during study, resulting in a sense of fluency which would lead to over-estimation of how much learning is occurring (Bjork et al., 2013). Additionally, a preference for blocked study may reflect habitual biases (Bjork et al., 2013; Pyc & Dunlosky, 2010) or a desire to avoid the greater effort associated with interleaving (Son & Simon, 2012).

Another factor that has been shown to modulate the benefits of interleaved or blocked study in laboratory contexts is the similarity relations between successive examples. The question is whether successive examples should maximize similarity or variation. In general, high similarity between examples facilitates identification of both similarities and differences through comparison (Gentner & Markman, 1994), and so could increase the benefits of both blocked and interleaved study sequences. However, high variability between examples can promote generalization (e.g, Braithwaite & Goldstone, 2012; Gómez, 2002), and so could increase transfer of learning to novel cases following study. Thus, there are competing reasons to expect benefits from the use of both successive similar and varied examples. Currently, little is known as to how learners might choose to regulate similarity or variation between successive examples, nor how such learner regulation would affect learning outcomes.

In summary, while existing research has provided considerable insight regarding the relative effectiveness of different study sequences in the laboratory, little is known about how learners behave in more ecologically valid contexts. Study behavior that is maladaptive in the laboratory may be more effective in naturalistic situations because of, for example, greater interest or relevance of the study situation. Having these possibilities in mind, the primary goal of the present study is to investigate learner choices regarding study sequencing during inductive learning, and associations of these choices with learning outcomes in an ecologically valid situation. A secondary goal is to investigate how much study variation students prefer during learning and how this interacts with study sequence.

Experiment

Method

In this experiment, students enrolled in an introductory psychology course were given practice calculating measures of central tendency using an online tutorial completed as homework following in-class instruction. On each tutorial trial, participants could choose which of three categories — mean, median, and mode — to study, and also could choose whether or not to vary the trial content (background story and data) with respect to the previous trial. Measures of central tendency were included on a subsequent mid-term exam which counted towards course grades, so participants were likely to be motivated to learn from the tutorial. Exam scores were matched to records of tutorial usage to identify associations between study behavior and exam performance. Importantly, the tutorial and its content were part of the normal class activities and the outcome measures included in the exam contributed to the students overall grade in the class.

Participants Undergraduate students enrolled in one of five sections of introductory psychology at Indiana University participated in this study. All students enrolled were required to complete the tutorial and exam as part of normal class activities. However, only data from students who consented for their data to be analyzed, and completed all parts of the study, were analyzed. The final sample consisted of 671 students.

Materials Two sets of four multiple choice questions each were constructed, one of which served as pretest and the other as posttest. Each set consisted of two procedural questions requiring exact calculation of measures of central tendency and two conceptual questions requiring qualitative inferences about these measures (see Table 1).

Table 1: Items used during pre- and posttest.

Procedural Questions	Conceptual Questions
Five cars were given safety ratings by consumer reports. Their ratings were: Spitfire = 3, Bentley = 7, Stanza = 8, Colt = 3, Lexus = 4. What are the mode, median and mean for this data set?	Imagine a difficult math test on which 13 students do very poorly, each getting a score of 1, 2 or 3 out of 100 possible points. However, the remaining 3 students get excellent scores: 96, 98, and 99. Will the mean be less than or more than the mode?
Three children in a family have ages of 7, 12, and 8. What are mean and median ages in this family?	There are 9 offensive players on a particular football team. On a particular game, the median number of yards gained by each player was 7 and no two players gained the same number of yards. If the worst and best performing offensive players are not considered, what will be the median of the remaining 7 players' gained yards?
Five pizzas were given quality scores by an expert taster. Their scores were: Pizza World = 8, Slices! = 3, Pisa Pizza = 2, Pizza a go-go = 4, Crusty's = 8. What are the mode, median and mean for this data set?	Imagine a vocabulary test in which 15 students do very well, getting scores of 98, 99, and 100 out of 100 possible points. However, the remaining 3 students get very poor scores: 5, 8, and 9. Will the mode be less than or more than the mean?
Three children in a family have shoe sizes of 5, 10, and 9. What are mean and median for the shoes sizes in this family?	There are 7 players on a particular basketball team. On a particular game, the median number of points scored by each player was 12 and no two players scored the same number of points. If the lowest and highest scoring players are not considered, what will be the median of the remaining 5 players' scores?

Thirty-two brief stories were designed for use as tutorial examples. Each story described a situation and presented a small dataset. For example, one story was "Several fishermen went fishing on the same day. Below you can find how many fish the different fishermen caught." The different stories used varied in context and details. The data set associated with each story was created online when the story was presented according to the constraints of the story and the students' selections (see below for details).

Students completed the pretest and tutorial online using an online platform created for this purpose.

Procedure During regular class sessions, students were assigned the pretest and tutorial as homework for class credit. They were also instructed that this homework assignment would be useful preparation for a future exam, which would include questions about measures of central tendency. Upon accessing the homework, participants first completed the pretest questions one at a time, without feedback. The tutorial began immediately after completion of the pretest. Participants first read a tutorial review describing how to calculate mean, median, and mode, followed by an explanation of the trial interface, followed by the tutorial trials.

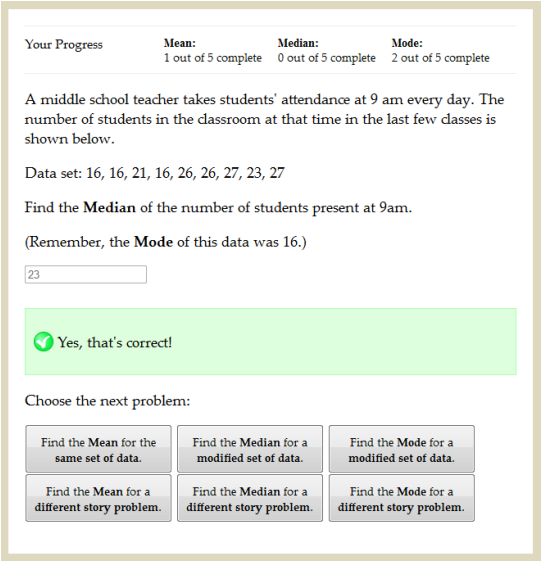

Figure 1: Tutorial interface for one of the trials during study. This example shows a problem and response feedback. The buttons at the bottom include all possible choices for the next problem.

Each tutorial trial presented one story along with a dataset, and requested participants to calculate the mean, median, or mode of the dataset (Figure 1). For the first trial, the category was always the mean, the story was selected randomly, and the dataset was generated quasi-randomly within the range specified for the story. The categories for subsequent trials (i.e. mean, median, or mode) were chosen by the participants. Thus, participants could choose to block study by studying each category several times successively, to interleave by cycling through the three categories repeatedly, or to adopt an intermediate approach.

Participants could also determine the degree of similarity or variation between successive trials by choosing whether each subsequent trial would involve the same or a new story. If a new story was chosen, a new dataset was generated quasi-randomly. If the same story was chosen, the dataset for the next trial was either identical to the current dataset if the category for the next trial had not been probed yet with that dataset, or a modified version of the current dataset otherwise.

As shown in Figure 1, the current story and dataset for each trial were displayed near the top of the screen, followed by an instruction to calculate mean, median, or mode. Once a response was submitted, feedback was displayed indicating whether the response was correct, followed by two rows of buttons, which were used by participants to choose what type of trial would appear next.

Trials involving the same or modified versions of the datasets used in the preceding trials also included prompts intended to facilitate comparison with the preceding trials. First, reminders of the answers to the preceding trials were displayed just above the response area. For example, the trial shown in Figure 1 requests the median, but includes a reminder of the answer to the previous trial, i.e. the mode. Second, trials involving modified datasets included descriptions of how the datasets were modified, with additions and deletions marked in the data.

The number of trials already completed for each category was displayed at the top of the screen. Participants were informed that they had to complete at least 5 trials for each category (regardless of correctness of responses). After this criterion was met, an option to end the tutorial became available. However, participants could continue the tutorial as long as they wished. Participants were encouraged to use a calculator, and a link to an online calculator was provided.

The four posttest questions were inserted into the standard mid-term exam for the course along with the remaining questions. This exam was administered using paper and pencil during class sessions, at least two weeks after the homework was made available. All students were required to take this exam, regardless of whether they had done the homework.

Results

Study behavior On average, participants completed 5.6 trials for mean, 5.5 for median, and 5.6 for mode, slightly more than the minimum required (5) in each case. Average accuracy during the tutorial was 86.0% for mean, 88.8% for median, and 97.0% for mode.Table 2 shows summary statistics for several aspects of participants' study behavior. First, for each participant, each transition between trials was classified as a category (or story) repetition if the category (story) of the current trial was also chosen for the next trial. Category and story repetition rates were then calculated by dividing the number of repetitions by the number of

transitions. Second, successive trials involving the same category (story) were grouped into blocks, and the number and average length of category (story) blocks were calculated. Finally, average spacing between category repetitions was calculated as the average number of different-category trials intervening between each trial and the next same-category trial. (Spacing between story repetitions was not calculated because a given story could not be repeated after a different story was chosen.)

Table 2: Summary Statistics for Study Behavior.

Variable	Mean *(SD)*	Min.	1st Quartile	Median	3rd Quartile	Max.
Number of trials	16.7 *(1.5)*	15.0	16.0	16.0	17.0	27.0
Category						
Repetition rate	63.3% *(30.0%)*	0.0%	44.8%	76.5%	86.7%	90.5%
Number of blocks	6.7 *(4.4)*	3.0	3.0	5.0	9.0	21.0
Avg. block length	3.6 *(1.9)*	1.0	1.7	3.6	5.3	7.3
Avg. spacing	1.8 *(0.8)*	1.0	1.0	1.7	2.5	3.0
Story						
Repetition rate	75.5% *(26.5%)*	0.0%	60.0%	85.7%	100.0%	100.0%
Number of blocks	4.9 *(4.2)*	1.0	1.0	4.0	7.0	19.0
Avg. block length	7.5 *(6.2)*	1.0	2.3	5.0	15.0	22.0

Participants showed clear tendencies to block study by category and to choose similar rather than varied trial content. To illustrate this tendency, we focus on category and story repetition rates; the other indicators of study yield similar results. The average category repetition rate was 63.3%, significantly higher than the rate of 33.3% which would result from random choice, $t(670)=25.86$, $p<.001$, and 78.5% of participants repeated categories at rates higher than chance. Similarly, the average story repetition rate was 75.5%, significantly higher than the rate of 50.0% which would result from random choice, $t(670)=24.90$, $p<.001$, and 82.7% of participants repeated stories at rates higher than chance. However, most participants (70.0%) switched stories at least once, although the tutorial could be completed without ever switching stories. Category and story repetition rates were uncorrelated, $r=.020$, $t(669)=0.509$, $p=.611$.

To investigate the possibility that participants might adapt study behavior based on perceptions of learning, category and story repetition rates were calculated separately for transitions following correct and incorrect responses for each participant, excluding those who gave no incorrect (32.3%) or no correct (0.0%) responses before any transition. Category repetition rates were lower following correct (66.0%) than incorrect (74.2%) responses, $t(453)=4.70$, $p<.001$. Likewise, story repetition rates were lower following correct (74.3%) than incorrect (79.4%) responses, $t(453)=3.44$, $p<.001$.

Test Performance Average accuracy was high on both the pretest (71.3%) and the posttest (84.6%), but was significantly higher on the posttest than on the pretest, $t(670)=14.11$, $p<.001$.

To investigate possible effects of study behavior on test performance, the posttest accuracy data were submitted to a linear regression with pretest score, category repetition rate, story repetition rate, number of tutorial trials, and tutorial accuracy as predictors. The model was significant, accounting for 13.0% of the variance in posttest score, $F(5,665)=19.83$, $p<.001$. The coefficients and significance of the various predictors are displayed in Table 3.

Table 3: Results of Regression Analysis of Posttest Accuracy. Asterisks indicate $p<.05$

Predictor	β	*F*	*p*	Sig.
Intercept	0.774	53.236	<.001	*
Pretest Score	0.293	89.683	<.001	*
Category Repetition Rate	0.064	6.695	0.010	*
Story Repetition Rate	0.006	0.044	0.834	
Accuracy During Tutorial	-0.026	0.165	0.685	
Number of Tutorial Trials	-0.009	3.530	0.061	

Not surprisingly, participants who scored well on pretest also scored well on posttest, as indicated by a significant effect of pretest score, uniquely accounting for 11.7% of the variance in posttest score, $\beta=0.293$, $F(1,665)=89.68$, $p<.001$. Moreover, a significant positive effect of category repetition rate was also found, uniquely accounting for 0.88% of the variance in posttest score, $\beta=0.064$, $F(1,665)=6.70$, $p=.010$, indicating that participants with higher category repetition rates tended to score higher on posttest. The effects of the other predictors were not significant, ps$>.05$. Analogous results were obtained when average category block length or average spacing between repetitions of the same category were entered into the model instead of average category repetition rate. A similar pattern of results was also found when analyzing separately the results for conceptual and procedural questions.

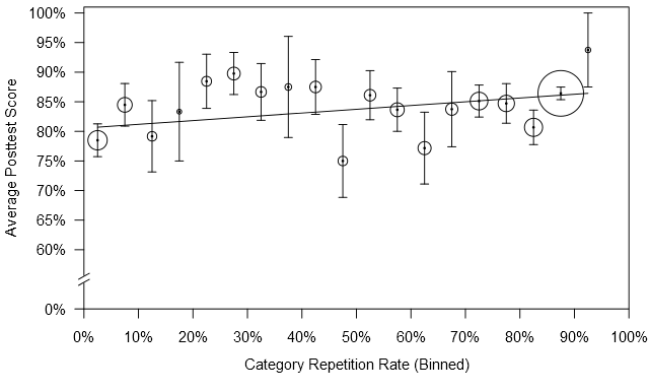

Figure 2: Average Posttest Score by Category Repetition Rate. Data binned by category repetition rate. The number of participants in each bin is represented by the area of the circles surrounding the data points (see text for details). Error bars represent standard errors of the mean.

The effect of category repetition rate on posttest score is illustrated in Figure 2. This effect is not evident in a traditional scatterplot because only five different posttest scores were possible for individual participants. Thus, participants were divided into bins, and average posttest scores within each bin were plotted. Each point in Figure 2 lies at the center of a 5%-wide interval of category repetition rates, and represents the average posttest score

among participants whose category repetition rates fell in that interval. The number of participants in each bin is represented by the area of the circles surrounding the data points. The regression line assumes average values for all predictors other than category repetition rate. The binned data were used for visualization only, and played no role in the above regression analyses.

Discussion

The main goal of this research was to investigate learner choices regarding how to sequence their study. The high rate of category repetition during the tutorial indicates that students have a tendency to block study by category. This result replicates the findings of Tauber et al. (2013), and extends them to an ecologically valid context, i.e. students studying in preparation for a test. As we mentioned in the Introduction, while this preference could reflect a belief that blocked study leads to better learning (Bjork et al., 2013; Son & Simon, 2012), it could also reflect an habitual bias towards blocked study (Pyc & Dunlosky, 2010) or an avoidance of extra work (Son & Simon, 2012).

We also find that participants predominantly chose to study similar successive examples, even across different concepts. Story repetition may have been preferred because it led to more fluid processing during study and thus increased perceptions of learning. Alternatively, participants may have wished to compare successive examples, and found this easier to do when successive examples involved similar content (Gentner & Markman, 1994), both between and within categories. However, we found no evidence that story repetition resulted in improved learning, as either of these beliefs might suggest.

Furthermore, category and story repetition rates were higher following incorrect than correct responses. This result may reflect metacognitive influences on study regulation, i.e. students might have perceived the current category as less well learned following incorrect than correct responses and chosen to repeat both categories and stories more often in the former case. Thus, our results dovetail well with previous findings that learners tend to defer study of well learned items, and to immediately repeat study of poorly learned ones (Pyc & Dunlosky, 2010; Son, 2004, 2010). Students may have preferentially repeated categories perceived to be far from a target level of mastery (Dunlosky & Hertzog, 1998), or alternatively, avoided repetition when it was expected to yield little incremental benefit (Metcalfe, 2009).

Our analyses also show an association between high category repetition rates and higher posttest scores. These results contrast with previous studies, which have found superior learning from interleaved study (Birnbaum et al., 2012; Kornell & Bjork, 2008; Rohrer & Taylor, 2007; Taylor & Rohrer, 2010; Wahlheim et al., 2011). Our data do not demonstrate that this association was causal, however, because category repetition rate was determined by the students rather than experimentally manipulated. Still, the effect of category repetition rate was significant even after accounting for effects of pretest score, accuracy during the tutorial, and number of tutorial trials. Additionally, category repetition rate was uncorrelated with pretest score, r=-.027, p=.491. These facts argue against an explanation of the effect of category repetition rate in terms of differences in pre-existing ability or diligence during the tutorial.

One possible explanation for the discrepancy between our findings and previous research relates to the tasks employed to assess learning. Interleaved study has led to superior performance on assessments requiring category discrimination (e.g. Kornell & Bjork, 2008). The posttest problems in the present study did not require category discrimination, because the categories involved in each problem were given explicitly in the problem statements. For such problems, sensitivity to internal category structure may be more useful than sensitivity to discriminative features, and blocked study has been argued to promote this type of learning more than does interleaved study (Carvalho & Goldstone, 2014; Goldstone, 1996). This interpretation supports the view, mentioned in the Introduction, that blocked and interleaved study may each be optimal for different tasks, in this case different testing situations.

An additional factor that may have favored blocked study in the present study relates to learner engagement. It is possible that the effects of blocked study differ depending on whether such study is chosen by the learner or by the teacher or tutoring system. For instance, a commonly mentioned drawback of blocked study is its repetitive nature, which might result in attention attenuation (Wahlheim et al., 2011). However, in the present study, participants' controlled the sequence of study during the tutorial and the tutorial questions are directly relevant for their course grades. Both these factors may have increased their engagement compared to previous studies in which study sequence was determined by the experimenter.

Another possibility is that the relative effectiveness of blocked study is increased when learners can choose strategically when to block. For example, by blocking specifically when it allows the student to test different theories and predictions, something that is not possible when the sequence is not under the student's control. An important next step would be to compare performance in situations in which students choose the study sequence and situations in which equivalent study sequences are presented to students who do not have control over the sequence of study.

Finally, from a methodological point of view, this experiment can serve as a model for research on pedagogically relevant issues. Unlike many laboratory studies, the present study has considerable ecological validity because it involved content belonging to the regular curriculum of a university course, and both intervention and assessment were tightly integrated with that course. In contrast to many classroom studies, online distribution of the tutorial allowed considerable control over the intervention and precise recording of the students' behavior for inclusion in subsequent analyses. We believe that similar

"in vivo" yet individually-controlled studies of learning in educational contexts (Koedinger, Aleven, Roll, & Baker, 2009) represent a major potential growth area for cognitive science.

Acknowledgments

This research was in part supported by National Science Foundation REESE grant 0910218, and Institute of Education Sciences, US Department of Education Grant R305A1100060 to RLG. PFC was also supported by Graduate Training Fellowship SFRH/BD/78083/2011 from the Portuguese Foundation for Science and Technology (FCT), and JRD by a National Science Foundation Graduate Research Fellowship under Grant No. DGE-1342962. We thank the instructors of participating sections of Introductory Psychology courses at Indiana University, Jim Craig, Melissa Wilkinson, Benjamin Sklar, and Susan Jones, for their cooperation and assistance in administering the research, and Tiffany Drasich for her assistance with data collection and processing.

References

Birnbaum, M. S., Kornell, N., Bjork, E. L., & Bjork, R. A. (2012). Why interleaving enhances inductive learning: The roles of discrimination and retrieval. Memory & Cognition, 41(3), 392–402. doi:10.3758/s13421-012-0272-7

Bjork, R. A., Dunlosky, J., & Kornell, N. (2013). Self-regulated learning: Beliefs, techniques, and illusions. *Annual review of psychology*, *64*, 417–44.

Braithwaite, D. W., & Goldstone, R. L. (2012). Inducing mathematical concepts from specific examples: The role of schema-level variation. In N. Miyake, D. Peebles, & Ri. P. Cooper (Eds.), *Proceedings of the 34th Annual Conference of the Cognitive Science Society* (pp. 138–143). Austin, TX: Cognitive Science Society.

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

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

Dunlosky, J., & Hertzog, C. (1998). Training programs to improve learning in later adulthood: Helping older adults educate themselves. In D. J. Hacker, J. Dunlosky, & A. C. Graesser (Eds.), *Metacognition in educational theory and practice. The educational psychology series.* (pp. 249–275). Mahwah, NJ: Lawrence Erlbaum Associates.

Elio, R., & Anderson, J. R. (1984). The effects of information order and learning mode on schema abstraction. *Memory & Cognition*, *12*(1), 20–30.

Gentner, D., & Markman, A. B. (1994). Structural alignment in comparison: No difference without similarity. *Psychological Science*, *5*(3), 152–158. doi:10.1111/j.1467-9280.1994.tb00652.x

Goldstone, R. L. (1996). Isolated and interrelated concepts. *Memory & Cognition*, *24*(5), 608–28. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/8870531

Gómez, R. (2002). Variability and detection of invariant structure. *Psychological Science*, *13*(5), 431–437.

Gureckis, T. M., & Markant, D. B. (2012). Self-directed learning: A cognitive and computational perspective. *Perspectives on Psychological Science*, *7*(5), 464–481. doi:10.1177/1745691612454304

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

Koedinger, K. R., Aleven, V., Roll, I., & Baker, R. (2009). In vivo experiments on whether supporting metacognition in intelligent tutoring systems yields robust learning. In *Handbook of metacognition in education* (pp. 897–964).

Kornell, N., & Bjork, R. A. (2008). Learning concepts and categories: Is spacing the "enemy of induction"? *Psychological Science*, *19*(6), 585–92. doi:10.1111/j.1467-9280.2008.02127.x

Kurtz, K. H., & Hovland, C. I. (1956). Concept learning with differing sequences of instances. *Journal of Experimental Psychology*, *51*(4), 239–243.

Medin, D., & Bettger, J. (1994). Presentation order and recognition of categorically related examples. *Psychonomic Bulletin & Review*, *1*(2), 250–254.

Metcalfe, J. (2009). Metacognitive judgments and control of study. *Current Directions in Psychological Science*, *18*(3), 159–163. doi:10.1111/j.1467-8721.2009.01628.x

Pyc, M. A., & Dunlosky, J. (2010). Toward an understanding of students' allocation of study time: why do they decide to mass or space their practice? *Memory & Cognition*, *38*(4), 431–40. doi:10.3758/MC.38.4.431

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

Son, L. K., & Simon, D. A. (2012). Distributed learning: Data, metacognition, and educational implications. *Educational Psychology Review*, *24*(3), 379–399. doi:10.1007/s10648-012-9206-y

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), 356–63. doi:10.3758/s13423-012-0319-6

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

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), 750–63. doi:10.3758/s13421-010-0063-y

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