The most efficient sequence of study depends on the type of test

Paulo F. Carvalho (pcarvalh@andrew.cmu.edu)
Human-Computer Interaction Institute, Carnegie Mellon University, 5000 Forbes Ave
Pittsburgh, PA 15213 USA

Robert L. Goldstone (rgoldsto@indiana.edu)
Department of Psychological and Brain Sciences & Cognitive Science Program, Indiana University, 1100 E. 10th St.
Bloomington, IN 47405 USA

Abstract

Previous research has shown that the sequence in which concepts are studied changes how well they are learned. In a series of experiments featuring naturalistic concepts (psychology concepts) and naïve learners, we extend previous research by showing that the sequence of study changes the representation the learner creates of the study materials. Interleaved study leads to the creation of relatively interrelated concepts that are represented by contrast to each other and based on discriminating properties. Blocked study, instead, leads to the creation of relatively isolated concepts that are represented in terms of their central and characteristic properties. The relative benefits of these representations depend on whether the test of conceptual knowledge requires contrastive or characteristic information. These results argue for the integrated investigation of the benefits of different sequences of study as depending on the characteristics of the study and testing situation as a whole.

Keywords: study sequence; interleaving; interrelated concepts;

Introduction

The sequence of study while learning concepts changes what is learned and how well it is learned. Therefore, it is perhaps not surprising that understanding how students should organize their study to promote learning has emerged as a major area of active interest in educational and cognitive science research. Previous research has focused on how different sequences might improve learning (Birnbaum et al., 2013), and how the benefits of different sequences might interact with different study conditions (Carvalho & Goldstone, 2015), materials (Carvalho & Goldstone, 2014), individual characteristics (Sana et al., 2016), or self-regulation (Carvalho et al., 2016).

When hard-to-discriminate concepts are studied in an interleaved fashion, by alternating the study of the different concepts, learning is improved compared to when different concepts are studied in separate blocks (Kornell & Bjork, 2008). However, the benefit of interleaved study is not universal. For example, it has been shown that when studying concepts that have high within-category diversity in their properties (for example, the category mammal which includes bats, cows, and whales), studying each concept in separate blocks can result in better learning (Carvalho & Goldstone, 2014). This apparent inconsistency lead to the proposal that different sequences of study emphasize different properties of the studied materials and thus might be more appropriate for different types of concept learning tasks (Carvalho & Goldstone, 2014). The Sequential Attention Theory (Carvalho & Goldstone, 2015), proposes a mechanism through which attention and encoding during blocked study are progressively directed towards the similarities among successive items belonging to the same category whereas attention and encoding during interleaved study are progressively directed towards the differences between successive items belonging to different categories. Because of this influence on cognitive processing, Carvalho and Goldstone (2015) propose that the sequence of study can accelerate or delay learning, depending on whether the constraints created by the sequence of study match those of the encoding situation (e.g., interleaved study in situations critically hinging on the encoding of differences between concepts, such as the study of highly similar concepts), or mismatch it (e.g., blocked study in the same situations).

In this work, we aim to extend these results to demonstrate that different encoding experiences will result in different representations that will be more or less appropriate depending on the requirements of the testing situation. Our proposal is as follows: because different information is encoded with different sequences of study, different sequences of study potentiate different representations of what was studied. More specifically, encoding the differences between concepts through interleaved study will tend to lead to the creation of interrelated concepts whose representations are contrasted away from each other by emphasizing or exaggerating their distinctive characteristic relative to each other (Corneille et al., 2006; Goldstone, 1996). Conversely, blocking will tend to lead to encodings of the similarities within each concept that will, in turn, create relatively isolated, stand-alone, representations (Goldstone, 1996).

These different representations, once created, are suited for different uses. Although an interrelated representation of two concepts will be helpful in a new context in which discriminating the previously learned concepts is important, isolated representations of the same concepts may not be as useful. Conversely, an isolated representation of a concept will include more information about all the properties of that concept, whether or not they serve to distinguish it from the other learned concepts, making it ideal for situations in which these details are relevant, such as when the concept must be differentiated from other new concepts possessing new distinctive features.

Students often create flashcards as a study and self-testing tool (Hartwig & Dunlosky, 2012). These flashcards might

include a definition or an example of a concept on one side of the card and the correct response on the other. When studying using examples, students might choose to study all the cards from one concept in a block or to interleave cards from different concepts. One important question, then, is if different sequences of examples will influence students' performance for different types of tests – a question that, to the best of our knowledge, has not been addressed before. This is not only an important question at the theoretical level – to know the representational differences created by different sequences – but also at the practical level because changing the sequence of study materials is an easy and cheap intervention that might have substantial influences on learning outcomes (Dunlosky et al., 2013). In fact, previous researchers have emphatically advocated presenting information interleaved whenever possible, warning students about the perils of blocked study (e.g., Bjork, Dunlosky, & Kornell, 2013), and it has been suggested as an important factor of which all new instructors should be aware (Deans for Impact, 2015).

For this purpose, we developed two experiments in which learners studied concepts of psychology (e.g., "Hindsight bias"; Rawson et al., 2015) in one of the sequences and were then tested in different situations, similar to common study practices by students. Importantly, some of the tests required discrimination between different concepts (e.g., multiple-choice test), whereas others required an independent representation of each concept (e.g., writing a definition). We consider writing a definition to require an independent representation because these definitions can be expressed without referring to other learned concepts. For example, a participant could write a definition for "availability heuristic" without having learned or remembered any of the other presented concepts (Goldstone, 1996). We predict that for tests that emphasize isolated, independent knowledge of the properties of each concept, such as writing a definition, participants will perform better following blocked study. Conversely, for tests that require discriminating different concepts, i.e., those that involve choosing between several options, participants will perform better with interleaved study.

Experiment 1

Method

Table 1: Participant demographic characteristics for Experiments 1 and 2.

Pair	Exp. 1	Exp. 2
Mean Age (SD)	33 (10)	36 (11)
Gender (% Females)	45.5%	68%
Education (% Bachelor's or higher)	50%	64%
Age Learned English (SD)	0.04 (0.21)	0.21 (1.13)

Participants. A group of twenty-eight people were recruited through Amazon's Mechanical Turk (https://www.mturk.com/). Data from 6 participants were excluded from analyses because of possible compliance issues (see below for details). The demographic characteristics of participants in the overall sample are presented in Table 1.

Stimuli. We used a stimulus set of introductory concepts and examples created by Rawson et al. (2015). The stimuli included 10 concepts taught in Introductory Psychology and 10 example situations for each concept, collected from textbooks of Introductory Psychology. The concepts were divided into two groups by relatedness. Each group contained unrelated concepts only, whereas across groups pairs of related concepts existed (see Table 2). Relatedness of the concepts was judged by the authors by comparing the definitions of the concepts and confirmed by analyzing the pattern of errors in multiple-choice questions without feedback in a pilot study. Previous research looking at sequence of study using these materials used this concept grouping as well (Rawson et al., 2015).

Table 2: Groups of concepts used in Experiment 1 and Experiment 2. Each row includes a pair of related concepts. Columns contain only unrelated concepts.

Pair	Group A	Group B
1	Availability Heuristic	Representativeness heuristic
2	Door-in-the-face technique	Foot-in-the-door technique
3	Hindsight bias	Counterfactual thinking
4	Fundamental attribution error	Deindividuation
5	Mere exposure effect	Social facilitation

Design and Procedure. This Experiment had two conditions manipulated within-subject: Study Sequence (Blocked vs. Interleaved) and Type of Test (Multiple-Choice Test vs. Definition Match Test vs. Write Definitions Test).

The experiment had three phases: pretest, study and test. Participants completed one pretest, two study phases and two tests phases in the following order: Pretest – Study 1 – Test 1 – Study 2 – Test 2. The first and second study phases were the same in every aspect except for the sequence of study and the concepts studied. One study phase was interleaved and the other blocked (order counterbalanced across participants). A different group of to-be-learned concepts was used in each study phase. In the interleaved condition learners studied an example of each concept before studying the same concept again (e.g., ABCABC…). Conversely, in the blocked condition learners studied all examples of each concept before starting a new concept (e.g., AABBCC…). Moreover, the test phase only tested the concepts learned in the immediately preceding study phase. Between each study and test phase participants completed a distractor task by watching a 4-minute video on an unrelated topic and answering a question about that video.

During the pretest phase participants were told that they would be presented with several psychology concepts that they were asked to rate regarding their familiarity/knowledge. Participants were told that not

knowing the concepts was not an issue for the study, would not impact their eligibility or payment, and that they should be honest in their responses. On each trial, the name of a concept was presented and participants had to rate on a scale from 1 ("Not familiar at all") to 7 ("Very familiar") how familiar they were with that concept. Following each rating, participants were asked to provide an example of that concept, or enter "I don't know" if they did not know any. Participants completed the pretest for the ten to-be-studied concepts across both study phases.

Following the completion of the pretest, participants completed the study phase. During the study phase, participants studied examples of situations depicting each of five concepts, one at a time and were asked to choose the name of the concept they thought the example instantiated. Participants were given feedback after each response. During study, participants studied five examples of each of the five concepts.

During the test phase, participants completed three types of tests: Multiple-Choice, Writing Definitions and Match Definitions, always in that order. The Multiple-Choice test used the same procedure as the study phase with new examples and without feedback. In the Writing Definitions, test participants were shown the name of each of the concepts studied one at a time and asked to write the best definition possible for that concept, based on what they had learned in the previous study phase. In the Match Definitions test participants were presented with the textbook definition of each concept, one a time, and asked to identify what concept that definition belonged to by pressing the corresponding button on the screen. The order of trials within each of the tests was randomized across participants. None of the test phase tasks had any time limit.

Results and Discussion

Because the study was conducted online without experimenter supervision, we first inspected the data in order to identify potential compliance issues. For each participant, we calculated the median response time during both study phases. The sample's median response time to complete the study phase was 10.5 seconds per problem (max: 22.9 sec./problem; min: 0.73 sec./problem). We calculated the 10th and the 90th percentiles for the distribution of median response times, 3.3 sec./problem and 16 sec./problem respectively, and used these values as a measure of non-compliance in the task. Responding too fast (faster than the 10th percentile) is likely due to participants who are not reading the problems and just advancing through the experiment quickly; similarly, longer response times (above that of the 90th percentile) are likely due to potentially distracted participants. Six participants were identified based on this analysis and their data were excluded from further analyses.

All the analyses below are ANCOVA analyses including average pretest score and counterbalancing condition as covariates.

Pretest. To analyze the data from the pretest we calculated 25th, 50th and 75th percentiles of the ratings (see Table 3). As can be seen, most participants showed little or no knowledge of the to-be-studied concepts (mean of approximately 2 in a 1-7 scale). The provided examples further confirmed this interpretation.

Table 3: Pretest results for Experiments 1 and 2 (1-7 scale).

	25 Percentile	50 Percentile	75 Percentile	*M*	*SEM*
Exp 1	1.00	1.55	2.03	1.71	0.16
Exp 2	1.34	1.79	2.39	1.97	0.16

Study Phase. Mean performance during the blocked study phase was 72% (*SEM* = 5%), whereas during interleaved study it was 67% (*SEM* = 5%). This difference was not statistically significant, $F(1,20) = 1.95$, $p = .169$, $\eta^2_G = .012$.

Test Phase. Two trained coders, blind to condition assignment, rated as correct or incorrect each of the written definitions. These two coders agreed 87% of the time and inter-coder reliability was high, Cohen's Kappa = .725, $p < .0001$. Disagreements were resolved by a third coder, also blind to condition assignment of the responses.

Performance for the test phase is depicted in Figure 1. As can be seen in the figure, the type of tests varied in their level of difficulty, with participants performing better in the Definitions Match test and worse in the Write Definitions test, $F(2,42) = 17.62$, $p = < .0001$, $\eta^2_G = 0.151$. Although there was no overall main effect of study sequence $F(2,42) = 1.36$, $p = .256$, $\eta^2_G = .006$, there was a significant interaction between type of test and study sequence, $F(2, 42) = 5.26$, $p = .022$, $\eta^2_G = .022$.

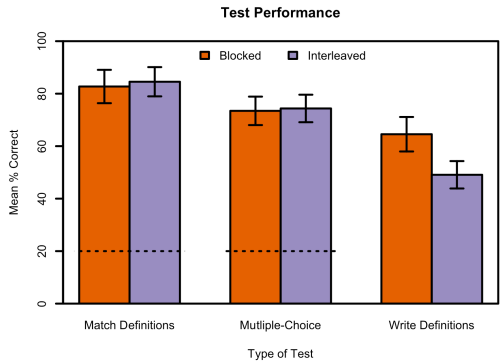

Figure 1: Results for the Test Phase of Experiment 1. Dotted lines represent chance level. Error bars represent standard errors of the mean.

To further investigate this interaction, we compared the effect of type of study sequence on each of the tests by calculating the difference in performance following blocked and interleaved study for each type of test (interleaved – blocked). The difference in performance between the two conditions varied across type of test, $F(2, 42) = 5.10$, $p = .011$, $\eta^2_G = .110$. Planned contrasts using FDR correction indicate that the effect of study sequence was significantly

different when comparing the Write Definitions test (M = -.15, SEM = .06) with the Multiple-Choice test (M = .009, SEM = .03), p = .033, and the Match Definitions test (M = .02, SEM = .05), p = .040, but not when comparing the Multiple Choice and the Match Definitions tests, p = .844.

These results are consistent with our proposal that blocked study encourages learners to develop independent, stand-alone representations rather than highlighting diagnostic features (i.e., those that discriminate between the concepts). Interleaved study emphasizes features that discriminate between concepts, which would be more helpful for a subsequent categorization task than a task that requires generation of stand-alone definition of the concept.

Experiment 2

Our main proposal in this paper is that blocked study creates relatively independent representations of each concept studied which emphasizes the concept's characteristic features. These independent representations include more details from each concept than what is fostered by the relatively interrelated representations created during interleaved study. In the context of studying examples of different concepts, we proposed that blocked study allows learners to more successfully write definitions of the concepts because a definition requires the type of knowledge that blocked study promotes; it is generally possible to write good definitions for the learned psychology concepts without mentioning other psychology concepts learned at the same time. Consistent with this hypothesis, Experiment 1 showed that following blocked rather than interleaved study, learners were more successful at writing definitions of concepts, but the groups did not differ on classifying examples.

However, when two concepts are highly related (e.g., foot-in-the-door and door-in-the-face technique) their definitions can be aptly construed in relation to each other. If they are studied together, one central feature to include in the definition is the feature that discriminates them. Thus, the fact that in the previous experiments learners studied in the same session concepts that were dissimilar from each other and varied in many properties (see Table 2) might have contributed to the pattern of results seen. Would studying similar concepts together change the pattern of results observed?

Studying related concepts together changes the learning task in at least three critical ways. First, studying similar concepts in the same session might result in the necessity to discriminate between similar situations in order to find the subtle differences between the two types of concepts. It has been shown before that the interrelated representations promoted by interleaved study are likely to improve learning in these situations of learning highly similar concepts (Carvalho & Goldstone, 2014). Second, the features that discriminate these related concepts are also characteristic features of the concept, unlike what is the case when the concepts are dissimilar (see Table 2). This means that interleaved study could promote representations appropriate for a writing definitions test through identification of differences between concepts, whereas these differences would not be likely to be highlighted in the previous experiment.

In sum, when similar items are studied in the same session, there are several reasons to believe that performance would benefit from interleaved study, even when the test requires learners to write definitions. However, when similar items are studied in separate sessions, as in Experiment 1, blocked study would promote best performance in a test requiring isolated representations, such as writing definitions.

To test this, we used a procedure similar to how students often organize their study. In most natural situations students are likely to randomly assign the topics to be studied to a study session or to follow the sequence of their textbook or instructor. Therefore, in this experiment we randomly assigned concepts to being studied either interleaved or blocked, instead of using different pre-defined groups of concepts that guarantee low between-category overlap as in the previous experiment. This results in a situation where similar concepts might be studied together or separately. We compare performance on multiple-choice and writing definitions tests following blocked or interleaved study in each one of these situations.

Method

Participants. A group of 36 people completed the experiment following recruitment through Amazon's Mechanical Turk (https://www.mturk.com/). Data from 3 participants were excluded due to self-reported previous participation in another study with the same materials. Data from an additional 8 participants were excluded from analyses because of possible compliance issues (see below for details). The final sample included 25 participants. Table 1 includes the demographic characteristics of participants in the overall sample.

Stimuli and Procedure. In this experiment, we used the same set of materials as in Experiment 1 and Experiment 2, but concepts were randomly assigned to be studied interleaved or blocked. Thus, in this experiment we did not force related concepts to be studied in separate phases.

The procedure was similar to the procedure used in Experiment 1 except for the following differences. Participants studied only eight concepts, four interleaved and four blocked. During study, participants saw four situations depicting each one of the concepts. After study, participants played a game of Tetris for 30 seconds.

The test phase included only a multiple-choice test and a writing definitions test, always presented in that order. During the multiple-choice test participants saw a total of four novel examples of the concepts studied, presented one at a time, and were asked to indicate which concept it illustrated.

Results and Discussion

We identified potentially non-compliant participants using the participants' response times during study. The sample's median response time to complete the study phase was 8.4

seconds per problem (max: 22.8 sec./problem; min: 0.47 sec./problem). The 10th and 90th percentiles for the distribution of median response times were 2.6 sec./problem and 16 sec./problem respectively. Eight participants were identified as outliers based on their falling outside of this range and their data were excluded from further analyses.

In all the analyses presented below, mean pretest score and counterbalancing condition were included as covariates.

Pretest. As in the previous experiments, participants showed little to no pre-training knowledge of the to-be-studied concepts (see Table 3).

Study Phase. Mean performance during the blocked study phase was 79% (*SEM* = 2.5%), while during interleaved study it was 73% (*SEM* = 4%). However, this difference was not statistically significant, F (1,25) = 2.65, p = .116, η^2_G = 0.04.

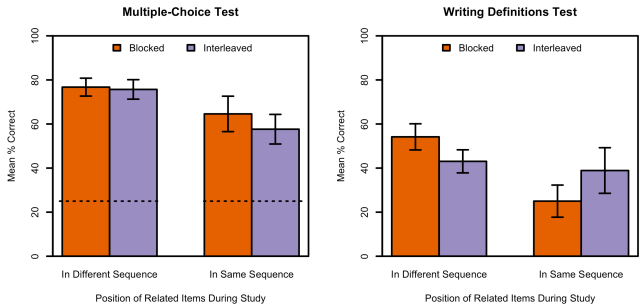

Figure 2: Results for the Test Phase of Experiment 2. Dotted lines represent chance level. Error bars represent standard errors of the mean.

Test Phase. Two trained coders, blind to condition assignment, rated as correct or incorrect each of the Written Definitions provided. The two coders agreed 84% of the time and inter-coder reliability was high, Cohen's Kappa = .611, p < .0001. Disagreements were resolved by a third coder, also blind to the condition assignment of the responses.

To analyze the results from the two tests used in this experiment we classified each concept based on whether it had been studied blocked or interleaved and whether its related concept (see Table 2) had been studied in the same sequence or in different sequences. When both related concepts were studied in the same phase and in the same sequence (e.g., "foot-in-the-door technique", "door-in-the-face technique" studied blocked), they were both classified as "Blocked" and "Same Sequence." However, when only one of the related concepts was studied, or the two related concepts were studied in different phases/sequences, both were marked "Different Sequences."

This classification of the concepts resulted in empty cells for participants who did not have both concepts studied in the Same Sequence and concepts studied in Different Sequences for both interleaved and blocked study. Because traditional repeated-measures ANOVA does not allow for the existence of empty cells and we wanted to maximize the inclusion of all data collected, here we used mixed model analyses and report Wald F tests and respective p-values using Kenward-Roger's approximation (Kenward & Roger, 1997). The results are depicted in Figure 2.

As we saw in the previous experiments, overall learners performed better on the Multiple-Choice test than when writing definitions, *Wald F* (1, 33.842) = 91.68, p < .0001. Similarly, the sequence of study had no overall effect on performance, *Wald F* (1, 33.948) < 1. No interaction was found between these two variables, *Wald F* (1,92.007) < 1.

However, the relatedness between concepts presented in the same sequence influenced performance. Overall, when participants studied the two related concepts in the same sequence their performance was lower (M = 46.52%, *SEM* = 1.50%) than when related concepts were not in the same sequence (M = 62.41%, *SEM* = 1.04%), *Wald F* (1, 24.284) = 15.60, p = .0006. Item relatedness also interacted with sequence of study and type of test, *Wald F* (1, 98.522) = 6.83, p = .010.

To further analyze this interaction, we explored the test results for each type of test separately. For the Multiple-Choice test, only the effect of relatedness reached statistical significance, *Wald F* (1, 24.264) = 9.43, p = .005. However, for the Writing Definitions test, in addition to a significant effect of item relatedness, *Wald F* (1, 25.649) = 9.81, p = .004, we also found a significant interaction between item relatedness and sequence of study, *Wald F* (1, 29.282) = 6.71, p = .015 (see right panel of Figure 5). As predicted by the results of Carvalho and Goldstone (2014), the relative relatedness between items modulates the relative benefit of each sequence for the Writing Definitions test. Moreover, consistent with the results of Experiment 1, we see that when similar items are not studied in the same sequence, performance in the Write Definitions test benefits from blocked study, although this effect was only marginally significant, t (35) = 1.90, p = .066, d = 0.317.

General Discussion

Overall, the results presented here show that the different sequences of study affect performance differently for different types of test. Studying examples of different concepts in a blocked sequence improves performance in a test requiring learners to provide a definition of the concept studied, whereas for other tests there is no difference in performance between the two sequences of study.

Consistently with previous research, we have argued that this pattern of results is related to the acquisition of different knowledge with each sequence. Whereas blocked study results in the creation of a relatively isolated representations (i.e., a stand-alone, independent representation of each concept), interleaved study results in interrelated representations (i.e., focusing on how a concept differs from other(s) studied at the same time; Corneille et al., 2006; Goldstone, 1996). Going one step further, these different representations are likely to be the result of differences in the underlying attentional and encoding processes (Carvalho & Goldstone 2015). The information attended to and encoded during study will dictate what type of representation is brought to a new situation and therefore what is available at test.

Moreover, we also saw that the effect of study sequence is modulated by whether discrimination based on subtle differences is necessary or not during study or test, such as is the case with the related concepts presented together in Experiment 2. We argued that this is the result of the pressures of the study and testing situation: when studying related concepts, interleaved study (and the interrelated representations it promotes) helps learners determine what discriminates between closely related concepts. This interpretation is consistent with the results of Carvalho and Goldstone (2014) showing that when learners studied similar categories, interleaved study improved learning, whereas when studying dissimilar categories, blocked study improved learning. Although the sample sizes used in the studies reported here might seem small, it is important to note that all critical comparisons were within-subject manipulations which increases the analytic power and that the effect sizes reported here are large and in line with previous similar research.

In sum, the two main contributions of the present work are as follow; first, it goes beyond existing demonstrations that blocking is better/worse than interleaving by showing how sequence affects what is learned by creating different representations given the same content. Second, it provides evidence for the context-dependent nature of learning and how the benefits of each sequence depend on the learning situation. This evidence adds to previous demonstrations that the best sequence of study depends on the type of material being studied (Carvalho & Goldstone, 2014; Patel et al., 2016), the type of study task (Carvalho & Goldstone, 2015; Rawson *et al.*, 2015), and whether students actively decide how to organize their study (Carvalho et al., 2016). These results also show the importance of developing *theories* of *why* one intervention is better than another. We have proposed a theory based on the similarities of the materials being learned and the nature of the task. When concepts are similar to each other, learners prioritize learning discriminating features. Writing definitions generally benefits from stand-alone representations unless the concepts being defined are similar to each other and benefit by being contrasted. The study of how an intervention interacts with the learning situation, we would argue, has the potential to not only provide a fuller understanding of how learning takes place, but also provide richer, more precise, recommendations for practice (Jonassen, 1982).

Acknowledgments

We are grateful to Katherine Rawson for sharing the stimuli set with us. Dustin Finch, Kaley Liang, Ashton Moody, Alifya Saify, and Shivani Vasudeva assisted with response coding. Work supported by NSF grant # 0910218 and IES grant # R305A1100060 to RG, and Fellowship # SFRH/BD/78083/2011 from the Portuguese Foundation for Science and Technology to PC.

References

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.

Bjork, R.A., Dunlosky, J., & Kornell, N. (2013). Self-regulated learning: beliefs, techniques, and illusions. *Annual Review of Psychology, 64*, 417–444.

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.

Carvalho, P.F., & Goldstone, R.L. (2015). The benefits of interleaved and blocked study: Different tasks benefit from different schedules of study. *Psychonomic Bulletin & Review, 22*(1), 281–288.

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

Carvalho, P.F., Braithwaite, D.W., De Leeuw, J.R., Motz, B.A., & Goldstone, R.L. (2016). An in vivo study of self-regulated study sequencing in introductory psychology courses. *PLoS ONE, 11*(3).

Corneille, O., Goldstone, R.L., Queller, S., & Potter, T. (2006). Asymmetries in categorization, cerceptual discrimination, and visual search for reference and nonreference exemplars. *Memory & Cognition, 34*(3), 556–567.

Deans for Impact. (2015). *The Science of Learning*. Austin, TX: Deans for Impact.

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.

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

Hartwig, M.K., & Dunlosky, J. (2012). Study strategies of college students: Are self-testing and scheduling related to achievement? *Psychonomic Bulletin & Review, 19*(1), 126–134

Jonassen, D.H. (1982). Aptitude-Versus Content Treatment Interactions: Implications for Instructional Design. *Journal of Instructional Development, 5*(4), 15–27.

Kenward, M.G., & Roger, J.H. (1997). Small sample inference for fixed effects from restricted maximum likelihood. *Biometrics, 53*(3), 983–97.

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

Patel, R., Liu, R., & Koedinger, K. (2013). When to Block versus Interleave Practice? Evidence Against Teaching Fraction Addition before Fraction Multiplication. In A. Papafragou, D. Grodner, D. Mirman, & J. C. Trueswell (Eds.), *Proceedings of the 38th Annual Conference of the Cognitive Science Society* (pp. 2069–2074). Austin, TX: Cognitive Science Society.

Rawson, K.A., Thomas, R.C., & Jacoby, L.L. (2015). The Power of Examples: Illustrative Examples Enhance Conceptual Learning of Declarative Concepts. *Educational Psychology Review, 27*(3), 483–504.

Sana, F., Yan, V.X., & Kim, J.A. (2016). Study Sequence Matters for the Inductive Learning of Cognitive Concepts. *Journal of Educational Psychology*.