Journal of Experimental Psychology: Learning, Memory, and Cognition

The Sequence of Study Changes What Information Is Attended to, Encoded, and Remembered During Category Learning

Paulo F. Carvalho and Robert L. Goldstone

Online First Publication, March 23, 2017. http://dx.doi.org/10.1037/xlm0000406

Journal of Experimental Psychology:
Learning, Memory, and Cognition

0278-7393/17/$12.00 http://dx.doi.org/10.1037/xlm0000406

The Sequence of Study Changes What Information Is Attended to, Encoded, and Remembered During Category Learning

Paulo F. Carvalho
Carnegie Mellon University

Robert L. Goldstone
Indiana University

The sequence of study influences how we learn. Previous research has identified different sequences as potentially beneficial for learning in different contexts and with different materials. Here we investigate the mechanisms involved in inductive category learning that give rise to these sequencing effects. Across 3 experiments we show evidence that the sequence of study changes what information learners attend to during learning, what is encoded from the materials studied and, consequently, what is remembered from study. Interleaved study (alternating between presentation of 2 categories) leads to an attentional focus on properties that differ between successive items, leading to relatively better encoding and memory for item properties that discriminate between categories. Conversely, when learners study each category in a separate block (blocked study), learners encode relatively more strongly the characteristic features of the items, which may be the result of a strong attentional focus on sequential similarities. These results provide support for the sequential attention theory proposing that inductive category learning takes place through a process of sequential comparisons between the current and previous items. Different sequences of items change how attention is deployed depending on this basic process. Which sequence results in better or worse learning depends on the match between what is encoded and what is required at test.

Keywords: interleaving, category learning, inductive learning, cue and category validity, comparison

There is a wide array of evidence that the sequence in which information is presented influences how we perceive, represent and learn new information (Bloom & Shuell, 1981; Brady, 2008; Clapper, 2014; Corcoran, Epstude, Damisch, & Mussweiler, 2011; Elio & Anderson, 1984; Helsdingen, van Gog, & van Merriënboer, 2011; Jones & Sieck, 2003; Li, Cohen, & Koedinger, 2013; Lipsitt, 1961; Mack & Palmeri, 2015; McDaniel, Fadler, & Pashler, 2013; Qian & Aslin, 2014; Samuels, 1969; Sandhofer & Doumas, 2008; Zeithamova & Maddox, 2009; Zotov, Jones, & Mewhort, 2011). In inductive category learning, for example, considerable research has analyzed the consequential differences between interleaving examples of different categories versus studying the same set of examples blocked by category. Although there is currently a sizable amount of research documenting the differences between these two sequencing strategies (Birnbaum, Kornell, Bjork, & Bjork, 2013; Kornell & Bjork, 2008; Kornell, Castel, Eich, & Bjork, 2010; Rohrer, 2009, 2012), and how they promote learning in different contexts (Carvalho & Goldstone, 2014b, 2015a; Kost, Carvalho, & Goldstone, 2015; Sana, Yan, & Kim, 2017; Zulkiply & Burt, 2013; Zulkiply, McLean, Burt, & Bath, 2012), less is known about the underlying cognitive processes leading to different learning from interleaved versus blocked study (Carvalho & Goldstone, 2015b). The main goal of this article is not, as with much of the previous research on interleaving and blocking of to-be-learned categories, to focus on which sequence is more beneficial but rather to identify the cognitive mechanisms that bring about empirically observed differences.

Most research on this topic finds robust differences in learning between interleaved and blocked study as measured by performance on a subsequent test task. The nature of this difference, however, is not uncontroversial. Whereas some research shows that interleaved study leads to improved learning and performance compared to blocked study (e.g., Rohrer & Taylor, 2007), other research shows the opposite (e.g., Carpenter & Mueller, 2013). For

Paulo F. Carvalho, Human-Computer Interaction Institute, Carnegie Mellon University; Robert L. Goldstone, Department of Psychological and Brain Sciences, and Program in Cognitive Science, Indiana University.

This work is part of a doctoral dissertation submitted by Paulo F. Carvalho to Indiana University. Portions of this work were presented at the 56th annual meeting of the Psychonomics Society in Chicago, IL and the 57th annual meeting of the Psychonomic Socieity in Boston, MA.

This work was supported by National Science Foundation (grant 0910218 to Robert L. Goldstone); Department of Education (IES grant R305A1100060 to Robert L. Goldstone); Portuguese Foundation for Science and Technology (Graduate Training Fellowship grant SFRH/BD/78083/2011 to Paulo F. Carvalho).

We thank the members of the Percepts and Concepts Lab for discussion, Rob Nosofsky, Rich Shiffrin, and Linda Smith for feedback and suggestions, and Dan Kennedy for valuable suggestions on the eye tracking analyses. Dustin Finch, Alanna Gilbert, Kaley Liang, Nurul Said, and Kristopher Shipman assisted with data collection. Dustin Finch, Abigail Kost, Alifya Saify, and Ashton Moody assisted with stimuli creation. The stimuli were designed by Xinrui Song.

The data from the studies presented in this article, as well as data from three additional replications of Experiment 1, are publicly available from: https://osf.io/2n8gy. The stimuli used are also available from the same repository.

Correspondence concerning this article should be addressed to Paulo F. Carvalho, Human-Computer Interaction Institute, Carnegie Mellon University, Newell-Simon Hall, 5000 Forbes Avenue, Pittsburgh, PA 15213. E-mail: pcarvalh@andrew.cmu.edu

example, Kornell and Bjork (2008) showed that learners who studied paintings from 12 different artists interleaved, rather than blocked, by artist were better at categorizing novel items at test (Experiments 1a and 1b; see also Kornell et al., 2010) and identifying whether the style of new paintings matched that of a previously studied artist or not (Experiment 2). Similar effects have been found using different materials and procedures (Carvalho & Goldstone, 2014b; Kang & Pashler, 2012; Li et al., 2013; Taylor & Rohrer, 2010; Zulkiply et al., 2012).

Conversely, Carpenter and Mueller (2013), showed that non-French speakers learned orthographic-to-phonological mappings in French (i.e., "-eau" and the corresponding sound /o/ in the words "bateau", "carreau", and "corbeau", and "-er" and the sound /e/ in the words "adosser", "attraper", and "baver") better when they studied different words with the same mapping blocked (bateau, carreu, corbeau, adosser, attraper, baver . . .), rather than interleaving words with different mappings (bateau, adosser, carreau, attraper, corbeau, baver . . .). This result has also been replicated with different materials and tasks (Carvalho & Albuquerque, 2012; Carvalho & Goldstone, 2011, 2014b; de Zilva & Mitchell, 2012; Goldstone, 1996; Kurtz & Hovland, 1956; Rawson, Thomas, & Jacoby, 2015; Zulkiply & Burt, 2013).

Carvalho and Goldstone (2014a, 2014b, 2015a, 2015b) proposed that this pattern of results can be parsimoniously explained by taking into account the requirements of the learning task and the attentional and encoding changes that each sequence of study promotes. The sequential attention theory (SAT; Carvalho & Goldstone, 2015b) proposes that during category learning, learners compare the current item with the previously studied one and, depending on the category assignment of the previous and current items, attend to similarities or differences between the two items. During each learning moment (e.g., a trial in a laboratory task), the learner evaluates similarities and differences between the current stimulus and the recollection they have of the previous item(s), as well as the correct category assignment of the previous exemplar and the current one. If the previous and current items belong to the same category, attention will be directed toward their similarities. However, if they belong to different categories, attention will be directed toward their differences. Across time, attention will be increasingly shifted toward relevant within-category similarities and between-category differences. This will, in turn, affect category representations, which will affect categorization decisions and recollection. With each new learning moment, the relevant properties will be progressively better encoded whereas irrelevant ones will be poorly or not encoded at all.

When categories are studied interleaved, the number of transitions between objects of different categories is frequent, which will result in attending to differences between categories on most trials by the process described above. In the same way, when categories are studied blocked, the likelihood of a within-category transition is high, which will increase attention toward within-category similarities by the same process (see Figure 1 for a schematic representation of this proposal). Furthermore, this process can also lead to encoding information that might not be central for learning the categories. For example, blocked study would lead participants to encode similarities within items of the same category that are also present in the other category and, therefore, cannot discriminate between the two categories.

Evidence for this process comes from studies using different types of categories. Carvalho and Goldstone (2014b) presented learners with low similarity or high similarity categories in either a blocked or interleaved sequence. Interleaved study improved categorization of novel items for high similarity categories (where identifying and encoding the differences between categories was key to strong categorization performance), whereas blocked study improved categorization of novel items for low similarity categories (where identifying and encoding the similarities among items of the same category was the major source of difficulty). Similar results have also been found using different materials (Zulkiply & Burt, 2013) and procedures (Carvalho & Goldstone, 2015a; Kost et al., 2015; Rawson et al., 2015), providing further support for SAT.

Although SAT provides a parsimonious theoretical framework and can account for a wide range of previous results showing interleaved and blocked study benefits (see Carvalho & Goldstone, 2015b), there is currently no direct evidence that learners attend to, remember, and encode different information in different sequences. The present work aims to fill this gap. To this end, we used stimuli composed of features that varied as to their category and cue validities.

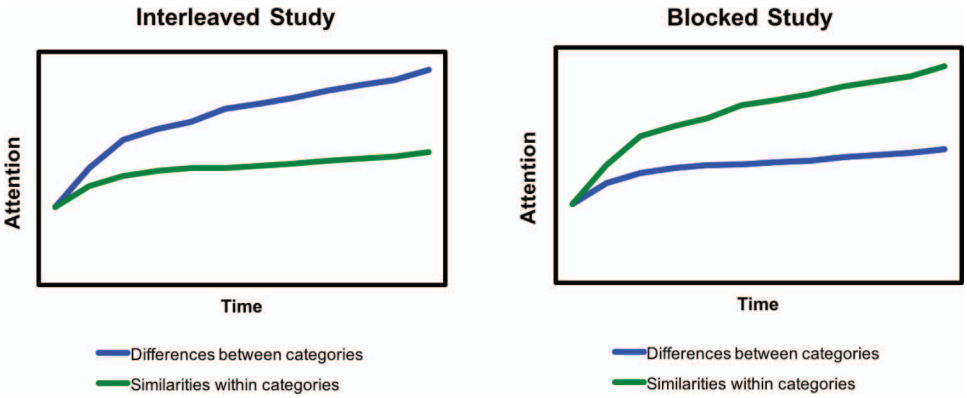

Figure 1. Schematic representation of the learning process proposed by the sequential attention theory and how interleaved and blocked sequences change what is encoded during study. See the online article for the color version of this figure.

The cue and category validity of a given feature are two measures of how that feature is distributed among all the objects present in the same and other categories. These measures are often times considered to be, respectively, measures of how diagnostic and characteristic a feature is for a category (Medin, 1983; Murphy, 1982; Murphy & Ross, 2005; Rosch & Mervis, 1975; Wisniewski, 1995), and they have been used in several models of categorization (e.g., Anderson, 1991; Murphy & Ross, 2005).

The cue validity of a feature *F* for a category *A* is defined as the probability of a category A given that the item has feature *F*, *P*("*X* is in Category *A*" | "*X* has feature *F*"). One way to calculate the cue validity of *F* for A is by dividing the total number of items in A that have F by the total number of items (across all categories) that have F. Higher values of cue validity are related with higher discriminability of that feature for a specific category. The category validity of feature F for category A is the probability that an item has feature F given that it is in Category A: *P*("*X* has feature *F*" | "*X* is in Category *A*"). Category validity can be calculated by dividing the total number of items in A that have feature F by the total number of items in A. Higher values of category validity indicate that the feature is very characteristic of that category (i.e., the feature is frequently present in a given category).

A feature that is strongly associated with a category would be both highly discriminative (high cue validity) and highly characteristic of that category (high category validity). For example, the feature *barks* is strongly associated with the category *dog* because most dogs bark (high category validity) and most things that bark are dogs (high cue validity). However, it is possible to achieve high cue validity with low category validity by, for example, decreasing the number of items that have that feature in all other categories. Similarly, it is possible to achieve high category validity with low cue validity by increasing the number of items in other categories that have the feature.

If, as SAT suggests, interleaving focuses attention toward the differences between successive items of different categories, a direct prediction of SAT is that during interleaved study learners will be relatively more likely to encode discriminative features (features with high cue validity), regardless of how characteristic of the category they are (i.e., how frequently they are present in the category, their category validity). This is because these features are more likely to change between successive items of different categories. Conversely, during blocked study learners will be relatively more likely to encode the characteristic properties of the category, the features which are frequently associated with the category label and thus have high category validity, regardless of their discriminative value (i.e., even when they have low cue validity). This is because these features are likely to be sequential similarities between items of the same category and SAT suggests that blocked study biases attention toward the similarities among successive items of the same category.

We tested these predictions in a series of studies. Participants in all experiments studied two categories interleaved and two categories blocked in two different phases. Following each study phase, participants completed one or more tests designed to assess feature encoding. In Experiment 1 participants completed a generalization task with novel items differing from studied items on critical features as well as a feature prediction task, designed to test differences in relative category significance of each feature following different sequences. In Experiment 2 we tested learners' memory for each type of feature across different retention intervals. Finally, in Experiment 3 we used eye tracker methods to identify which features learners looked at most during study.

Experiment 1

To study which features are encoded during interleaved and blocked study we manipulated the statistics of the features in the training items such that some features had high cue validity but low category validity (discriminative features), whereas other features had low cue validity but high category validity (characteristic features; see Table 1). To test our prediction of differential encoding of these features following study with different sequences (see Introduction), we created transfer items that differed from studied items on the characteristic features only and items that did not differ from studied items on either of these two types of features. We predicted that following interleaved study, participants will be equally good at classifying both types of items because both types of items retain the features more likely to be encoded during study—the discriminative features. Conversely, following blocked study, participants will be better at classifying items that do not differ on the characteristic features compared to items with novel characteristic features. Even though both items can be easily classified into one of the categories based on their discriminative features, because during blocked study the similarities within categories are more likely to be encoded, characteristic features will have a greater tendency to be encoded as relevant for categorization. Their absence at test is expected to adversely impact performance. We should note that this is not to say that learners

Table 1
Category Structure for the Stimuli Used in Experiments 1, 2, and 3

		Dimension				
Category	Item	1	2	3	4	5
A	1	2 (1, .3)	1 (.5, .7)	1 (.5, .7)	1 (.5, .7)	4 (.5, .3)
A	2	2 (1, .3)	1 (.5, .7)	1 (.5, .7)	1 (.5, .7)	5 (.5, .3)
A	3	2 (1, .3)	2 (1, .3)	1 (.5, .7)	1 (.5, .7)	3 (.5, .3)
A	4	1 (.5, .7)	2 (1, .3)	1 (.5, .7)	1 (.5, .7)	3 (.5, .3)
A	5	1 (.5, .7)	2 (1, .3)	2 (1, .3)	1 (.5, .7)	5 (.5, .3)
A	6	1 (.5, .7)	1 (.5, .7)	2 (1, .3)	1 (.5, .7)	4 (.5, .3)
A	7	1 (.5, .7)	1 (.5, .7)	2 (1, .3)	2 (1, .3)	4 (.5, .3)
A	8	1 (.5, .7)	1 (.5, .7)	1 (.5, .7)	2 (1, .3)	3 (.5, .3)
A	9	1 (.5, .7)	1 (.5, .7)	1 (.5, .7)	2 (1, .3)	5 (.5, .3)
B	1	3 (1, .3)	1 (.5, .7)	1 (.5, .7)	1 (.5, .7)	4 (.5, .3)
B	2	3 (1, .3)	1 (.5, .7)	1 (.5, .7)	1 (.5, .7)	5 (.5, .3)
B	3	3 (1, .3)	3 (1, .3)	1 (.5, .7)	1 (.5, .7)	3 (.5, .3)
B	4	1 (.5, .7)	3 (1, .3)	1 (.5, .7)	1 (.5, .7)	3 (.5, .3)
B	5	1 (.5, .7)	3 (1, .3)	3 (1, .3)	1 (.5, .7)	5 (.5, .3)
B	6	1 (.5, .7)	1 (.5, .7)	3 (1, .3)	1 (.5, .7)	4 (.5, .3)
B	7	1 (.5, .7)	1 (.5, .7)	3 (1, .3)	3 (1, .3)	4 (.5, .3)
B	8	1 (.5, .7)	1 (.5, .7)	1 (.5, .7)	3 (1, .3)	3 (.5, .3)
B	9	1 (.5, .7)	1 (.5, .7)	1 (.5, .7)	3 (1, .3)	5 (.5, .3)

Note. Numbers represent a specific feature value on each dimension. They represent independent feature values across Dimensions (i.e., a 2 on Dimension 1 is unrelated to a 2 on Dimension 2). A value of 2 or 3 is always a discriminative feature, whereas a value of 1 is a characteristic feature. Which part (eyes, legs, arms, antenna, mouth) corresponded to each dimension was counterbalanced across participants. Cue and category validity values for each feature value are presented in parentheses (cue validity, category validity).

will not attend to discriminative features during blocked study if these are relevant for the task at hand. Rather, our proposal, is that the characteristic features will be relatively better encoded in the blocked than interleaved study condition. In fact, because we used a categorization task, an overall degree of bias toward discriminative features might be expected (Markman & Ross, 2003; Yamauchi & Markman, 2000).

Moreover, we also predicted that following each study sequence, learners would rate the importance of different features for categorization differently. If learners encode the characteristic features relatively more effectively during blocked study, we should expect not only that novel items without the characteristic features would be harder to classify but also that learners would rate characteristic features as relevant for categorization to a greater extent than do participants who studied the categories interleaved. To test this, participants completed a feature prediction task in which they were asked to either rate features for how likely it would be that an item with that feature belongs to a specific category (cue validity test) or how likely it would be that an item in Category X would have that specific feature (category validity test).

Method

All experimental protocols and consent materials for this and subsequent studies were reviewed and approved by the Indiana University Institutional Review Board.

Participants. A total of 100 undergraduate students at Indiana University agreed to participate in this study in return for partial class credit. Participants were randomly assigned to either the cue validity test ($n = 53$) or the category validity test ($n = 47$) groups.

Apparatus and stimuli. The stimuli used were images of alien creatures (see left panel of Figure 2 for examples of one item from each of two species of one of the families). Two types (families) of aliens were created, each including two categories (species). The two families of alien creatures differed only on their visual appearance and not on the underlying properties of the feature space used to create them.

Each alien creature was composed of five feature-dimensions (i.e., Arms, Legs, Eyes, Mouth and Antenna), and each creature could have different feature values for each dimension. A total of 5 feature values were created for each dimension (see Figure 3 for some examples). During the study phase, participants studied items that followed the structure in Table 1 (this table also includes information about the cue and category validity of each feature value).

For Dimensions 1–4 there were three possible feature values. Two of these values predicted category membership (discriminative features; values 2 and 3 in Table 1). However, these discriminative features were overall infrequent in the space (each having 33% probability of being present, low category validity). The third value (Value 1 in Table 1) was a characteristic feature. This feature did not predict category assignment (low cue validity) but was highly frequent in both categories, therefore presenting high category validity for both categories. Dimension 5 could assume one of 3 feature values. These features did not predict category assignment and were overall infrequent in the space (random features, both low cue and category validity), and were included to increase variability in the space and the presence of unique items.

An additional transfer set was created for each category. This set was composed of characteristic-changed items and characteristic-preserved items (see right panel of Figure 2 for examples). Both types of items differed from studied items on the feature values presented on Dimension 5. The random features were replaced by novel features that participants had never seen before (see Table 2). In addition, characteristic-changed items also included a novel feature that replaced the characteristic features presented during study (see Table 2).

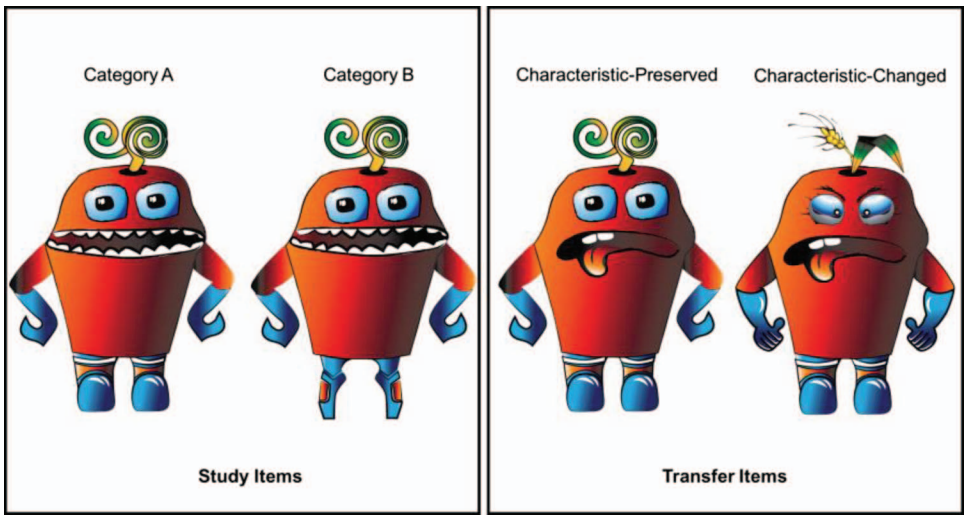

Figure 2. Example of stimuli from one of the families used in Experiments 1, 2 and 3. The left panel includes an example of each of the categories studied. The right panel includes an example of each of the novel items presented during the transfer task (both transfer items belong to Category A; equivalent items existed for Category B). See the online article for the color version of this figure.

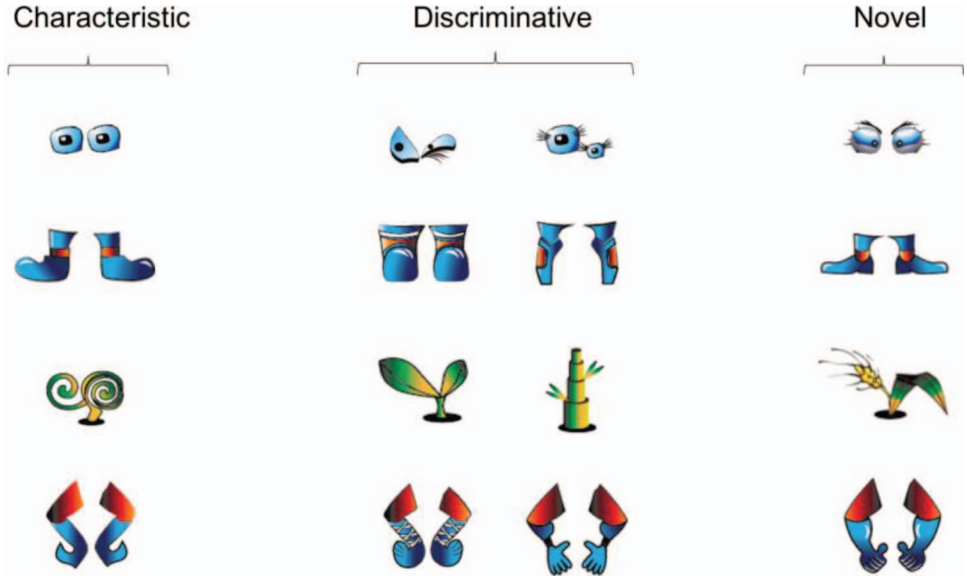

Figure 3. Example of one combination of features used in the stimuli of Experiments 1, 2, and 3. Which feature was chosen to be characteristic, discriminative, or novel was counterbalanced across participants. See the online article for the color version of this figure.

One family of alien creatures (two species) was randomly selected to be studied blocked and the other family (two species) was studied interleaved. The feature-values and dimensions used were counterbalanced across participants.

Each category was given a unique label. At the start of the experiment one label was randomly picked for each category from the following pool: beme, kipe, vune, coge, zade, and tyfe.

Table 2
Category Structure for the Stimuli of One of the Categories Used During the Transfer Task of Experiment 1

			Dimension				
Category	Type of item	Item	1	2	3	4	5
A	Characteristic-changed	1	2	4	4	4	6
A	Characteristic-changed	2	2	4	4	4	7
A	Characteristic-changed	3	2	2	4	4	6
A	Characteristic-changed	4	4	2	4	4	7
A	Characteristic-changed	5	4	2	2	4	6
A	Characteristic-changed	6	4	4	2	4	7
A	Characteristic-changed	7	4	4	2	2	6
A	Characteristic-changed	8	4	4	4	2	7
A	Characteristic-changed	9	4	4	4	2	6
A	Characteristic-preserved	10	2	1	1	1	6
A	Characteristic-preserved	11	2	1	1	1	7
A	Characteristic-preserved	12	2	2	1	1	6
A	Characteristic-preserved	13	1	2	1	1	7
A	Characteristic-preserved	14	1	2	2	1	6
A	Characteristic-preserved	15	1	1	2	1	7
A	Characteristic-preserved	16	1	1	2	2	6
A	Characteristic-preserved	17	1	1	1	2	7
A	Characteristic-preserved	18	1	1	1	2	6

Note. Numbers represent a specific feature. Numbers are independent across Dimensions (i.e., a 2 in Dimension 1 is not the same as a 2 in Dimension 2). Participants were also tested on comparable items for Category B. Which part (eyes, legs, arms, antenna, mouth) corresponded to each dimension was counterbalanced across participants.

Design and procedure. Participants studied four categories. Two categories were studied interleaved and another two were studied blocked. This experiment had three phases, always presented in the same order. Participants started by studying two categories (study phase), followed by classification of novel items (transfer task), and then a feature-prediction task in which participants rated the likelihood of different features being part of a category based on what they learned during the study phase. Participants completed the three phases with one pair of categories from one of the alien families in one of the study sequences and then completed the three phases again with a different pair of categories from a different alien family and a different study sequence.

During the study phase, participants were told that a new planet had been discovered and that new species of alien creatures had been found. Each creature could be classified into one species based solely on their visual appearance. Participants were then instructed to classify each creature presented as belonging to one of two species by pressing a button with the name of the species on the screen.

On each trial an image of an alien creature was presented in the center of the screen. After 1,500 ms,[1] two buttons were presented at the bottom of the screen and participants were tasked with providing a classification. Immediately after a classification, feedback was presented in the center of the screen by replacing the image classified with "CORRECT!" or "INCORRECT." There was no time limit for responses but participants were instructed to respond as quickly and accurately as possible. Participants completed a total of four blocks of study. In the interleaved condition, in all blocks, after the presentation of an item from one category an item from the other category was presented. In the blocked con-

[1] We included this delay to approximate the timing in this procedure to that of procedures used in our previous studies and to guarantee that participants studied the item before pressing the response key and advance the trial.

dition, two blocks contained only presentations of one category, followed by two blocks with presentations from the other category.

The transfer task was similar to the previous phase except participants were not given feedback following each response and the order of presentation of items belonging to each of the two categories was randomized. There were three types of items presented during the Transfer Phase. Items that participants had studied in the previous phase (studied items), items that differed from the studied items only in the values of the random features (characteristic-preserved items) and items that differed from the studied items in the values of the random features as well as the value of the characteristic features (characteristic-changed items).

During the feature prediction task, participants were shown images of different features. There were four types of features: features that were characteristic of both categories during study (characteristic features), features that discriminated between the two categories during study (discriminative features), features that were presented for the first time during the transfer task to replace the studied characteristic feature (novel-transfer features), and novel features never presented during the experiment (novel features).

Half the participants were shown the feature and asked to rate on a 1–100 scale the likelihood of an alien in a specific category having the feature presented. A rating of 1 was described as indicating that it would be unlikely, whereas a rating of 100 would indicate that it would certainly belong to that category. The other half of the participants were shown the feature and asked to rate on a 1–100 scale the likelihood of an alien with that feature belonging to a specific category. For both conditions, trials were presented randomly and participants were instructed to use the entire scale when providing their ratings. The protocols used in this and subsequent studies were approved by the Office of Research Compliance of Indiana University.

Results and Discussion

The two main questions of interest in this experiment are (a) Does the sequence of study change what features are encoded during study and therefore available at test, and (b) Does the sequence of study change the perceived relevance of different features for category assignment?

Transfer task. To answer the first question, we can look at accuracy during the transfer task.[2] If learners relatively strongly encode the similarities within each category during blocked study, as we have proposed, then transfer performance for characteristic-changed items should be worse than for characteristic-preserved items. Similarly, if during interleaved study learners are relatively more likely to encode the differences between the categories, they should show equivalent transfer performance for both types of items because both preserve the discriminative feature unaltered. Because the dependent variable of interest in this analysis (accuracy) is proportional in nature, accuracy data were first submitted to an empirical logit transformation. We plot the raw accuracy data.

Overall, the sequence of study did not affect performance on the transfer task, $F(1, 99) < 1$. The overall effect of type of item was also not statistically significant, $F(1, 99) = 2.99$, $p = .087$, $\eta_G^2 = .003$. However, as it can be seen from the plot depicting the results of the transfer task in Figure 4, there was an interaction between

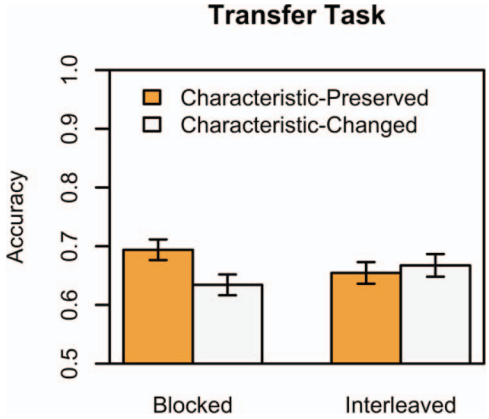

Figure 4. Results of the transfer task of Experiment 1. Chance performance in this task was 0.50. Error bars represent standard errors of the mean. See the online article for the color version of this figure.

type of item and study sequence, $F(1, 99) = 9.15$, $p = .003$, $\eta_G^2 = 0.01$. Post hoc t tests using Holm correction (Holm, 1979) further confirm that this interaction follows the predicted pattern. Participants who studied the items blocked classify new items that preserve the characteristic feature more accurately than new items that do not preserve this feature, $t(99) = 3.18$, $p = .008$, $d = 0.319$. Contrariwise, participants who studied the items interleaved classified both types of items with similar accuracy levels, $t(99) = 0.86$, $p = .523$, $d = 0.086$. Moreover, participants' sensitivity to the discriminative features did not vary between the two sequences of study; participants were equally good at classifying characteristic-changed items in both conditions, $t(99) = 2.10$, $p = .115$, $d = .210$.

The hypothesis outlined in the introduction assumes that following interleaved study, learners should show equivalent classification accuracy for all the transfer items because during study they encoded the discriminative features relatively efficiently and therefore they are more strongly associated with category assignment and better remembered than the other features. This implies that following interleaved study, learners should be as good at categorizing novel items with changed characteristic features as at categorizing studied items, because these items do not differ on the discriminative features that participants have encoded.

Conversely, during blocked study, our hypothesis was that learners encode more efficiently the similarities within the categories compared to interleaved learners. This includes to some degree the discriminating features but to a greater degree, the characteristic features. Therefore, during the transfer task we expected a decrement in performance for characteristic-changed items relative to studied items due to the absence of the encoded characteristic features, but not for characteristic-preserved items because these preserve most of the features that were encoded during study (both the characteristic and discriminative features).

[2] For this and following studies, we did not analyze performance during the study phase because the sequence of presentation of the categories was deterministic for both conditions (AAABBB and ABABAB), which led to ceiling performance across both conditions. This level of performance might not be indicative of learning but participants' understanding of the transition probabilities (see Carvalho & Goldstone, 2014b).

To test this prediction, we can compare performance during the transfer task between studied items and each of the novel items: characteristic-changed and characteristic-preserved. Both types of novel items differ from studied items by including novel features in Dimension 5 (noncharacteristic and nondiscriminative features). In addition, characteristic-changed items also include new features that replace the characteristic features studied. If performance for characteristic-changed items is worse than for studied items, this means that a change in the characteristic features negatively impacts performance. If performance is also worse for characteristic-preserved than Studied items, then any change from studied items negatively impacts performance. This comparison keeps constant the effect that time elapsed between study and the transfer task might have had, as we are comparing performance only for items presented during the transfer task. For this analysis we calculated, for each participant, a difference score between transfer performance for Studied and each of the novel items. A value of zero on this measure indicates that participants' performance at test is not influenced by introducing novelty. whereas negative values indicate a negative impact of the change. We are interested in differences between the conditions depending on what changed between studied and test items. The results of this analysis are plotted in Figure 5.

We found a significant interaction between sequence of study and type of change in the transfer items, $F(1, 99) = 7.95$, $p = .006$, $\eta^2_G = .009$. Pairwise t tests using Holm correction for multiple comparisons show that whether the characteristic feature was changed or not had an impact on performance for blocked study only, $t(99) = 3.189$, $p = .012$, $d = 0.319$. No other pairwise comparison reached statistical significance (all $ps > .191$). Overall, this pattern of results provides further evidence that replacing the characteristic features in characteristic-changed items has a negative impact on generalization following blocked study, whereas other novelty introduced at test does not affect performance, and that both sequences of study result in equal sensitivity to the discriminative features.

In sum, the results of the transfer task show that blocked study results in better encoding of the similarities among items of the same category, namely the characteristic features. Overall, learners in the interleaved sequence seem to ignore characteristic features, whereas during blocked study these features are attended to and effectively encoded.

Feature-prediction task. To answer the second question, that is, whether the sequence of study changes the perceived relevance of different features for category assignment, we can look at the results of the feature-prediction task. If the sequence of study influences the perceived relevance of different features for categorization, then we should see an interaction between the type of feature being rated and the sequence of study.

We asked participants to rate how predictive a feature was of a particular category (cue validity test group) or how predictive a category was of a particular feature (category validity test group). For both types of questions, there were two critical features presented: the characteristic features studied and the discriminative features studied. We also included two other features to serve as controls: the novel features introduced during the transfer phase to replace the characteristic features (transfer features) and novel features never seen before.

Overall participants rate as more predictive the studied features than any of the novel features (all $ps < .05$) for all the comparisons analyzed and rate discriminative features as more likely to belong to their correct category than the opposite category (all $ps < .05$). This pattern of results indicates that they understood the task and are in fact using their recollection of the feature distribution (see also Figure 6 for a full comparison of the ratings of several features). Considering now the critical features (discriminative and characteristic features), although participants provide overall higher mean ratings for cue validity test questions ($M = 48.39$, $SD = 22.02$) than for category validity test questions ($M = 40.52$, $SD = 25.60$; $F(1, 98) = 4.21$, $p = .043$, $\eta^2_G = 0.019$), the overall pattern of results is similar for both types of questions and the type of question did not interact with any of the other factors ($ps > .053$). Therefore, in all the analyses below we collapse across type of question.

Across both study sequences, participants provided higher ratings for discriminative ($M = 59.73$, $SD = 24.89$), than characteristic features ($M = 49.90$, $SD = 19.41$), $F(1, 99) = 43.13$, $p < .0001$, $\eta^2_G = 0.047$. This results suggests that learners are aware of

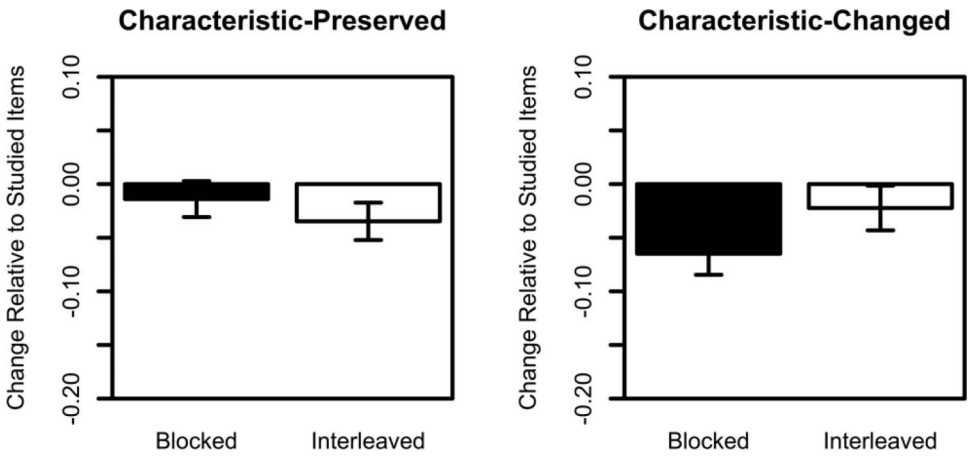

Figure 5. Effect of different changes introduced in the novel items of the transfer task compared to performance for studied items during the transfer task. Error bars represent standard errors of the mean.

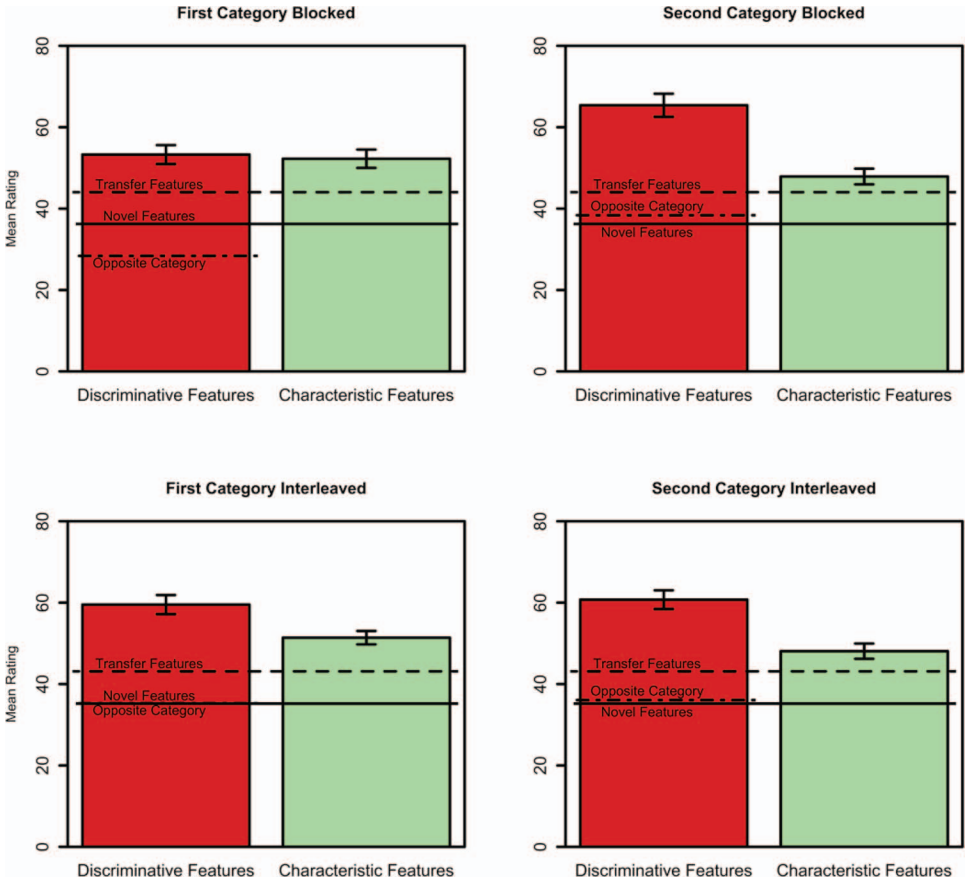

Figure 6. Results of the feature prediction task of Experiment 1. Transfer features refer to features that were not presented during the study phase but were presented as novel features in the transfer task. Novel features refer to features that had not been presented at any previous point during the experiment. Opposite category refers to ratings for that feature when probed about the opposite category. For interleaved study, the first category studied was defined as the category of the first stimulus studied by the participant, for each participant. Error bars represent standard errors of the mean. See the online article for the color version of this figure.

the high cue validity of the discriminative feature, which would be expected from successful learning. Moreover, they also overestimate its category validity compared to characteristic features. This could be the result of a bias to assume that discriminating features are also characteristic of a category.

However, no main effect of sequence, $F(1, 99) < 1$, or interaction between type of feature and sequence of study, $F(1, 99) < 1$, were found. Nonetheless, it is possible that because all participants learned the categories presented, regardless of sequence of study, the discriminative feature becomes more salient for all learners. One important difference might be how this takes place over time. For blocked study, because learners start by studying one category in isolation and encoding the similar properties among stimuli of the same category, it is unlikely that they overestimate the importance of the discriminative feature over the characteristic ones until they start studying the second category. For interleaved study, on the other hand, learners should encode more effectively differences between categories. Therefore, they should be aware of this critical difference from the beginning of the training and start encoding the discriminative feature as more relevant for both categories. Thus, we predict that for blocked study the ratings for the first and second category will differ, whereas for interleaved study the pattern will be the same both for the first and second category studied.

The plot in Figure 6 shows the mean ratings for defining and characteristic features for the first category studied (left panel) and for the second category studied (right panel), for blocked study (top row), as well as for interleaved study (bottom row).[3] As can be seen from Figure 6, there is a three-way interaction between type of feature to be rated (characteristic vs. discriminative), which category is being probed (first vs. second), and the sequence of study (interleaved vs. blocked), $F(1, 99) = 6.77$, $p = .011$, $\eta_G^2 = 0.005$. For participants who studied the categories blocked, the difference in ratings for discriminative and characteristic features depends on whether the first or second category is being probed, $F(1, 99) = 18.03$, $p < .0001$, $\eta_G^2 = 0.030$. More specifically there is no difference between ratings for discriminative and character-

[3] For interleaved study, the first category studied was defined as the category of the first stimulus studied by the participant, for each participant.

istic features for the first category studied, $t(99) = 0.38$, corrected $p = .702$, $d = .038$; but for the second category the ratings are higher for discriminative features than for characteristic features, $t(99) = 6.07$, corrected $p < .0001$, $d = 0.607$. Conversely, for interleaved study there is no interaction between the type of feature and what category is being probed on the ratings provided, $F(1, 99) = 3.23$, $p = .076$, $\eta^2_G = 0.060$, for both categories participants rate the discriminative feature as more relevant for categorization.

Taken together, the results of this experiment make two important contributions by (a) showing that different sequences of study result in the encoding of different properties of the stimuli and (b) that these differences are likely to result from a process of in time stimulus comparison. Participants who studied the categories blocked showed worse transfer to new items that differed from studied items on the characteristic (but nondiscriminative) feature when compared to both new items that did not differ on this feature and studied items. This is an impressive result because, to perform well in this task, participants had only to learn which discriminative features were associated with each category. The fact that, despite that, we find a role of characteristic features following blocked study speaks to how emphasized they were during study in that sequence. Moreover, these participants rated as equally relevant for categorization both types of features for the first category studied, but not for the second category studied. This suggests that attention toward and encoding of the discriminative feature is the result of having the opportunity to contrast two categories. Conversely, learners who studied the items interleaved show no decrement in performance for transfer items that vary the characteristic features and rate the discriminative features as more likely to occur than characteristic features regardless of which category is being probed.

Experiment 2

The main goal of this experiment was to extend the results of Experiment 1 to a memory task that allowed us to test the prediction that learners remember different information from different study sequences. This new test provides convergent evidence to the results of Experiment 1, showing that the sequence of study changes not only what is encoded, but this differential encoding has consequences for what information about the studied items learners will remember.

Participants completed the same study task as in Experiment 1. Following this study phase, learners completed a recognition memory task. During this task, participants were presented with features one at time and asked to rate on a scale from 1 to 6 whether they had seen that feature during study or not. We presented studied characteristic features as well as discriminative features for each category studied (first vs. second). We also included foils that resembled the studied features but varied on either color or shape. Different groups of learners completed the memory task at different intervals: immediately after study, 3 min after study, or 1 to 3 days after study. We included several retention intervals because it has been shown before in the literature on spacing of verbatim repetitions that the length of the retention interval has an effect on the spacing effect (Cepeda, Pashler, Vul, Wixted, & Rohrer, 2006; Crowder, 1976; Donovan & Radosevich, 1999; Janiszewski, Noel, & Sawyer, 2003; Murray, 1983), and similar results have been shown with interleaved study of category items (but see Carvalho & Goldstone, 2014a; Ste-Marie, Clark, Findlay, & Latimer, 2004).

We predict an interaction between the sequence of study and which features are remembered more effectively. Characteristic features should be relatively well remembered following blocked compared to interleaved study, whereas discriminative feature should be relatively well remembered following interleaved compared to blocked study. Moreover, because these memory differences are the direct result of how efficiently different features were encoded during study, we expect a similar decay function for all features across both sequences of study.

Method

Participants. A total of 302 undergraduate students at Indiana University agreed to participate in this study in return for partial class credit. Participants were randomly assigned to either the immediate ($n = 104$), the 3-min delay ($n = 74$), or the extended delay ($n = 124$) conditions. Participants in the extended delay condition were asked to return to the lab within 3 days of the initial study session for the test session. Participants were free to schedule their follow-up date for a time convenient to them within these constraints.

Data from 25 participants were excluded from analyses. Data from 2 participants in the Immediate condition were excluded due to a computer error during data collection. Data from 22 participants in the extended delay condition were excluded due to failure to complete the test session. Only one participant in the extended delay condition returned to the lab within 24 hours, therefore data from this participant were excluded from analyses.

Apparatus and stimuli. The stimuli used during the study phase of this experiment were the same as for Experiment 1. During the memory phase we used the characteristic and discriminative features studied as well as two types of novel, never studied features: features that varied in shape relative to each of the studied features, and features that varied in color relative to each of the studied features (see Figure 7 for some examples). These novel features were designed to be highly similar to the studied features, increasing the need for a precise recollection of the studied feature.

Design and procedure. There were two phases in this experiment: study phase and memory phase. In all conditions participants started by completing a study phase similar in every aspect to that of Experiment 1. Following the end of the study phase participants completed a memory task. Participants in the immediate test condition completed the memory task immediately after the end of the study phase, and participants in the 3-min delay condition completed a 3-min distractor task where they were asked to answer trivia questions. Participants in the extended delay condition completed only the study phases for each sequence of study in the first visit to the lab and were asked to return to the lab within 3 days to complete the memory phase (participants were never told that there would be a memory task, only that a follow-up was necessary).

The memory task was identical for all groups of participants. Participants were shown a feature in the center of the screen and asked to rate between 1 (*"sure never seen it"*) and 6 (*"sure seen it"*), whether they had seen that feature during the study phase. Studied features were presented on 1/3 of the trials; on the remaining trials, novel features were presented. features could be classi-

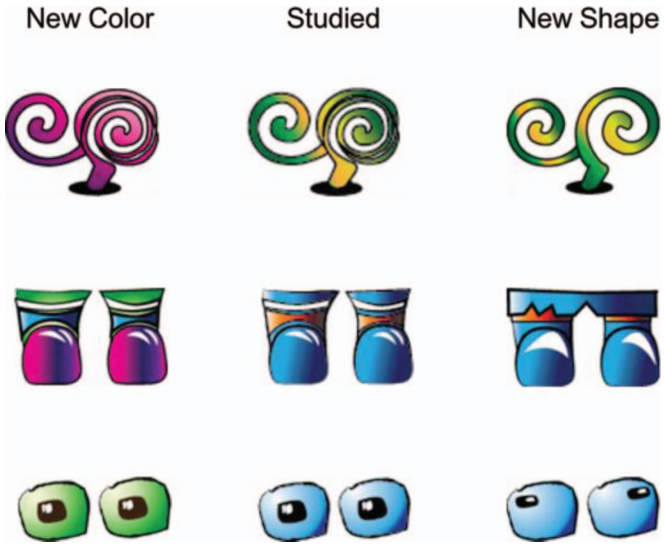

Figure 7. Examples of studied, new color, and new shape features presented during the memory task of Experiment 2. See the online article for the color version of this figure.

fied as characteristic, discriminative of the first category or discriminative of the second category, depending on whether that feature had been characteristic of both categories or discriminative of one of the categories during study (for studied items), or was a variation of a characteristic or discriminative features during study (for novel items). Trials were presented in random order and participants were instructed to use the whole scale when providing their ratings.

Results and Discussion

The final sample included a total of 102 participants who completed the memory test immediately after finishing the study, 74 participants who completed the memory task following a 3-min delay, 62 who completed the memory task 2 days after the study phase, and 39 who completed the memory task 3 days after the study phase. In all subsequent analyses, we will use these groups to establish a forgetting function for the different types of features studied and different study sequences.

We started by looking at the average ratings for each type of feature (characteristic vs. discriminating of the first category vs. discriminating of the second category), retention interval (immediate vs. 3 min vs. 2 days vs. 3 days), and sequence of study (interleaved vs. blocked). The average ratings are presented in Table 3. Participants' average ratings of the items decreased with increasing retention intervals, $F(1, 275) = 8.32$, $p = .004$, $\eta^2_G = 0.013$, and varied depending on the type of feature being rated, $F(2, 550) = 21.46$, $p < .0001$, $\eta^2_G = 0.012$. Importantly, participants' ratings of different types of features were modulated by the sequence of study, $F(2, 550) = 14.23$, $p < .0001$, $\eta^2_G = 0.007$. Following blocked study, participants provided on average similar

Table 3
Means and Standard Deviations of the Ratings for the Three Different Types of Features Presented During the Memory Task of Experiment 2

Retention interval	Characteristic	Discriminative first category	Discriminative second category	Average
Blocked study				
Immediate	3.626 (1.969)	3.324 (1.937)	3.504 (2.043)	3.484 (1.987)
3 min	3.507 (1.944)	3.269 (1.880)	3.584 (1.929)	3.453 (1.922)
2 days	3.477 (1.883)	3.289 (1.839)	3.435 (1.938)	3.401 (1.888)
3 days	3.35 (1.944)	3.051 (1.846)	3.316 (1.914)	3.239 (1.905)
Average	3.522 (1.941)	3.263 (1.889)	3.484 (1.973)	3.423 (1.938)
Interleaved study				
Immediate	3.565 (1.947)	3.45 (1.921)	3.422 (1.955)	3.479 (1.942)
3 min	3.497 (1.936)	3.489 (1.943)	3.389 (2.003)	3.458 (1.960)
2 days	3.571 (1.895)	3.435 (1.834)	3.347 (1.947)	3.451 (1.894)
3 days	3.357 (1.935)	3.308 (1.927)	3.286 (1.916)	3.317 (1.925)
Average	3.519 (1.931)	3.437 (1.909)	3.377 (1.960)	3.444 (1.934)
Grand Total	3.52 (1.936)	3.35 (1.900)	3.431 (1.967)	3.434 (1.936)

Note. Marginal averages are presented for each type of feature, each delay condition, and each Schedule × Type of Feature Combination.

ratings for characteristic and discriminative features of the second category, $t(276) = 0.92$, $p = .360$, $d = 0.055$. Ratings for both of those features were on average higher than for discriminative features of the first category, $t(276) = 5.73$, $p < .001$, $d = .344$ and $t(275) = 4.65$, $p < .0001$, $d = 0.279$, respectively. Following interleaved study, on the other hand, participants provided equivalent ratings for both discriminative features of the first and second categories, $t(276) = 1.49$, $p = .138$, $d = 0.089$, and between the characteristic and discriminative features of the first category $t(276) = 1.93$, $p = .108$, $d = 0.116$. Participants' ratings were higher for characteristic than discriminative features of the second category, $t(276) = 3.18$, $p = .005$, $d = 0.191$.

This pattern of results suggests that learners remember differently different properties of the studied items depending on the sequence in which the categories were studied. However, this measure is not specific to memory accuracy, but could instead or in addition represent changes in criterion used. That is, higher ratings could be related with the use of lower criteria for some features (differences between conditions resulting from overall higher ratings to both novel and old items) or to better memory accuracy for some features (differences resulting from overall higher ratings for old items and lower ratings for novel items). For a more precise estimate of learners' memory, we calculated a measure of memory accuracy and repeated the analyses above using that measure. Because participants provided a rating on a scale and not a binary response, we adopted a signal detection approach to derive this measure. More specifically, we used receiver operating characteristics (ROCs) analysis to derive recognition accuracy from a set of ratings.

A ROC curve relates the rate of correct classification responses (hits) and incorrect classifications (false alarms; FAs) for a variety of criteria. We calculated hits and FAs by participant for each type of item (characteristic feature vs. discriminative feature of the first category vs. discriminative feature of the second category), for five different criteria to classify an item as having been seen before: (a) a rating of 2 or greater, (b) a rating of 3 or greater, (c) a rating of 4 or greater, (d) a rating of 5 or greater, and (e) a rating of 6. To correct for high hit and FAs rates we added 0.5 to the numerator and 1 to the denominator of the calculation for all criteria (Stanislaw & Todorov, 1999). The aggregate ROC curves for blocked and interleaved study, each type of feature and retention interval are presented in Appendix A.

To quantify the shape of the ROCs, we represented the ROCs in z-space by taking the z-score (the inverse of the standard cumulative normal distribution with mean of 0 and standard deviation of 1) of the hit and FAs rates (the aggregate zROC curves are presented in Appendix B). zROC curves have several properties extensively studied in recognition memory experiments (Yonelinas & Parks, 2007), thus this approach allows us to investigate learners' memory by relying on these known properties of zROCs.

Two common measures derived from zROC curves are the slope and the intercept of the ROC curve in z-space (z-slope and z-intercept). The z-slope indicates the symmetry of the ROC curve. Although different models suggest different sources for asymmetries (z-slopes different from 0), it is usually associated with encoding variability of the source items, recollection/familiarity differences, or attentional factors (Yonelinas & Parks, 2007). Differences in z-intercept are associated with changes in sensitivity, that is, recognition accuracy (Yonelinas & Parks, 2007). Several studies have shown that factors known to increase recognition memory result in increased z-intercepts (e.g., increasing the number of study presentations, Ratcliff, Sheu, & Gronlund, 1992). Moreover, overall, this measure is independent of changes in the asymmetry of the ROC curves, often taken as evidence of a dual-process memory system (Yonelinas & Parks, 2007). Because we are interested in memory accuracy changes for different types of features between the two study sequences and the associated forgetting curve, in the following analyses we will focus only on z-intercepts. To avoid potential documented issues associated with analyzing only the averaged version of the ROCs (see, e.g., Malmberg & Xu, 2006; Morey, Pratte, & Rouder, 2008), we calculated for each participant the best-fitting line for the ROC curves in z-space using conventional regression methods and took the intercept of that line. For each participant, we took the intercept of the best fitting line for the 5 points provided (one for each possible criteria).

Figure 8 shows the mean z-intercepts for each type of item in each sequence of study and study-test delay condition. As one would expect, participants' recognition accuracy decreases with increasing temporal delay, $F(1, 275) = 13.72$, $p = .0003$, $\eta^2_G = 0.02$; however this decrease does not vary between the two sequences of study, $F(2, 550) = 2.53$, $p = .11$, $\eta^2_G = 0.002$. Moreover, overall, participants' recognition accuracy (z-intercept) was better following interleaved ($M = 0.633$, $SD = 0.632$) than blocked study ($M = 0.568$, $SD = 0.630$), $F(1, 275) = 4.06$, $p = .045$, $\eta^2_G = 0.003$. Finally, there was also an overall effect of type of item, $F(2, 550) = 13.46$, $p < .0001$, $\eta^2_G = .009$, with worse memory for the discriminative features of the first category ($M = 0.516$, $SD = 0.633$), compared to both the characteristic features ($M = 0.635$, $SD = 0.629$), $t(276) = 4.08$, $p < .0001$, $d = 0.245$, and the discriminative features of the second category ($M = 0.650$, $SD = 0.627$), $t(276) = 4.89$, $p < .0001$, $d = 0.294$. No difference was found between memory for the discriminative features of the second category and memory for the characteristic features, $t(276) = 0.53$, $p = .595$, $d = 0.032$.

Of particular relevance is the significant interaction between type of item and study sequence evident in in the plot of z-intercepts by study sequence and retention interval presented in Figure 8, $F(2, 550) = 6.11$, $p = .002$, $\eta^2_G = 0.004$. This interaction was not affected by changes in the duration of the retention interval, $F(1, 99) = 1.40$, $p = .247$, $\eta^2_G = 0.001$. Whereas following interleaved learners remember the different features to the same level of accuracy, $F(2, 552) < 1$, following blocked study the type of feature influences their memory accuracy, $F(2, 552) = 19.23$, $p < .0001$, $\eta^2_G = 0.0257$. To further describe this influence, we conducted a series of pairwise post hoc t tests using Holm correction for multiple comparisons. Following blocked study participants' memory is significantly more accurate for the characteristic features ($M = .635$, $SD = .640$) and the discriminative features of the second category ($M = 0.644$, $SD = 0.603$), compared to the discriminative features of the first category, ($M = .425$, $SD = .626$), $t(276) = 4.95$, $p < .0001$, $d = .297$, and $t(276) = 5.67$, $p < .0001$, $d = .340$, respectively. No difference in memory accuracy was found between the discriminative features of the second category and the characteristic features following blocked study, $t(276) = 0.22$, $p = .830$, $d = 0.013$.

Overall, the results of this experiment show that learners' memory of what is studied is affected by the sequence in which

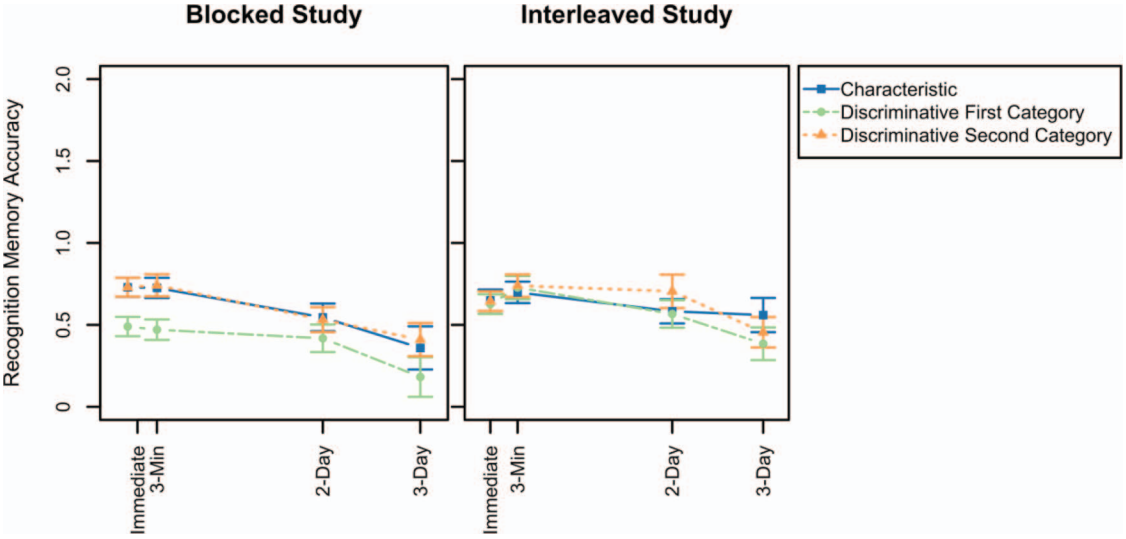

Figure 8. Results for the memory task of Experiment 2. The recognition memory accuracy measure used is the z-intercept of receiver operating characteristics curves. Error bars represent standard errors of the mean. See the online article for the color version of this figure.

information is studied. We saw an interaction between the type of feature and the sequence of study on the two measures studied: learners' raw ratings and memory accuracy. For blocked study, characteristic features were relatively better recalled than the discriminating features of the first category. This is consistent with the proposal that during blocked study learners encode more effectively similarities within the categories. Moreover, as we saw in Experiment 1, the fact that memory accuracy did not differ between characteristic and discriminating features of the second category suggests a temporal profile by which, after contrast with another category, discriminating features might gain some greater encoding. One interpretation of this pattern of results could be the existence of a recency effect for the second category studied blocked (Baddeley & Hitch, 1977; Glanzer, Adams, Iverson, & Kim, 1993; Glanzer & Cunitz, 1966). However, the fact that the duration of the retention interval did not significantly change this pattern of results argues against this possibility (Phillips & Christie, 1977). Moreover, a recency bias interpretation of these results would predict that characteristic features would be more strongly associated with the second category than the first, which was not the case in the feature prediction task of Experiment 1.

In contrast, following interleaved study, learners' memory accuracy did not differ considerably between the three types of features, even though ratings were slightly higher for the discriminating features of the second category studied. One possible reason why we saw no differences in memory accuracy for the different features studied interleaved might be connected with differences in number of presentations. characteristic features were presented more times than discriminating features; the fact that there is no difference between these two types of features can be interpreted as a powerful encoding of the discriminating features during interleaved study, despite having been studied considerably less times.

Overall, these results are consistent with our proposal that blocked learners more strongly encode the similarities within categories relative to interleaved learners, resulting in a relative benefit toward encoding the characteristic features of the categories, whereas interleaved learners more strongly encode the differences between items of different categories relative to blocked learners, resulting in a relative benefit toward the discriminating features of the categories.

Experiment 3

The previous two experiments suggest that learners encode different properties of the items being studied during interleaved compared to blocked study. Moreover, these encoding differences result in recognition memory differences at several time delays. Although this is compelling evidence that the sequence of study changes what is encoded, we do not have direct evidence of attentional changes during study that lead to encoding differences between the two sequences. SAT suggests that learners will learn to attend to different information because of the way examples are organized in the different sequences, which ultimately will result in different encoding and memory for the information studied (Carvalho & Goldstone, 2015b).

Previous research using eye tracker technology has demonstrated that eye gaze can be a good indicator of attentional flexibility. This measure of overt attention matches modeling predictions of how attention changes during category learning (Kruschke, Kappenman, & Hetrick, 2005; Rehder & Hoffman, 2005a, 2005b) and suggests that learners can rapidly learn to differentially allocate their attention (Blair, Watson, Walshe, & Maj, 2009). For example, Chen, Meier, Blair, Watson, and Wood (2013; see also Blair, Watson, & Meier, 2009) showed that learners were more likely to fixate category-relevant than category-irrelevant features of the objects and that learners' patterns of fixation show consistent regularities. Thus, eye gaze can be a diagnostic measure of overt attention during learning that has been

shown to map to predictions of attentional changes during category learning.

Building on this previous research, in this experiment we use eye tracking technology to study how the sequence of study changes the moment-by-moment allocation of attention. According to SAT we predict that learners will spend relatively more time looking at similarities between successive items during blocked study and differences between successive items during interleaved study.

Method

Participants. A total of 91 undergraduate students participated in this study in return for partial study credit. Data from 24 students were excluded due to computer issues ($n = 4$), experimenter error ($n = 2$), unsatisfactory calibration ($n = 3$), or insufficient number of valid eye tracking samples (more than 40% of samples in a trial with missing information) for more than 40% of the total trials in the study ($n = 15$).

Apparatus and stimuli. The stimuli used during the study phase of this experiment were the same as in Experiment 1. Images were presented on a 23" monitor with 1,920 × 1,080 pixels resolution. Eye gaze information was collected using an integrated Tobii Eye Tracker (model TX300; Tobii AB, Danderyd, Sweden), at a 120 Hz sampling frequency. Participants sat approximately 60 cm away from the center of screen. Their position was adjusted so that participants' eye gaze was centered on the screen (relative to both horizontal and vertical axes) at the beginning of the experiment.

Stimulus presentation, data response collection, and eye tracking data were collected using E-Prime 2.0 (Psychology Software Tools, Inc, Pittsburgh, PA) with Tobii Extensions (Tobii Technology BA, Stockholm, Sweden). Eye tracking data was analyzed using custom code written in R (R Core Team, 2013). Regions of interest (ROIs) were defined as a rectangular shape around each of the features of the item. For features that were divided (e.g., arms), two ROIs were defined but the average of the two was computed for comparison with the other ROIs.

Design and procedure. Participants started by completing a 9-point calibration of the eye tracking system. Following satisfactory calibration, participants started the experimental task. The procedure was similar to Experiment 1 except for the following changes. Participants completed only the study and transfer phases (one for interleaved and another for blocked study), but no feature prediction task. During the transfer task, participants saw only novel items. These changes were introduced to maintain the total duration of the study within 40 min.

Results and Discussion

The main hypothesis being tested in this experiment is that learners will attend to different properties of the items depending on the sequence of study. We started by analyzing looking time for the discriminative features and characteristic features. To this end, we calculated, for each study sequence and type of feature, the total time each participant spent looking at that feature during each of the study trials. To account for base-rate differences of how likely each feature was to occur, the total time looking during each trial was divided by the number of features of each type (discriminative or characteristic) presented on that trial. For analyses, the corrected total time looking at each type of feature was summed across trials for each participant and sequence to achieve a total looking time score.

The left panel of Figure 9 shows the total time participants spent looking at each type of feature depending on study sequence. During blocked study participants spend less time looking at all the features ($M = 923$, $SD = 676$), than during interleaved study ($M = 1{,}351$, $SD = 1{,}019$), $F(1, 66) = 11.32$, $p = .001$, $\eta^2_G = 0.050$.

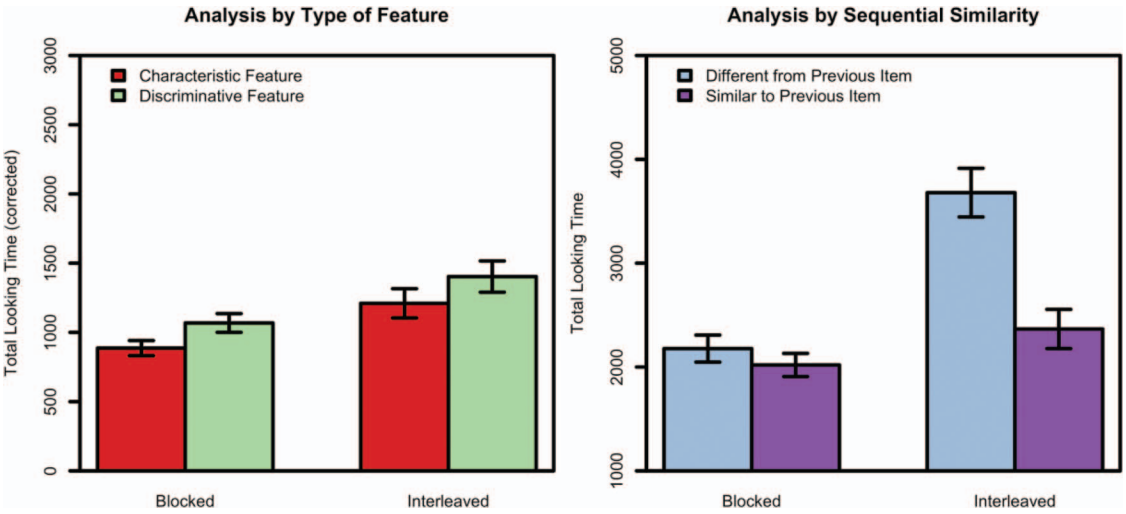

Figure 9. Results of looking time analyses for Experiment 3. The left panel includes the results for looking time analyses by type of feature (characteristic vs. discriminative). The right panel includes the results when analyzing looking time by sequential similarity (features that differed between trial N and $N - 1$, and those that did not vary). Error bars represent standard errors of the mean. See the online article for the color version of this figure.

Moreover, participants spend more time looking at the discriminative (M = 1,236, SD = 778) than the characteristic features (M = 1,048, SD = 705), $F(1, 66) = 45.00$, $p < .0001$, $\eta^2_G = 0.017$. There was no interaction between the sequence of study and the type of feature on looking time, $F(1, 66) < 1$.

These results indicate that there are no differences on how much learners attend to the two types of features presented depending on the sequence of study. At first look these data might seem contradictory to those of the previous experiments. If learners attend to the discriminative features more in both sequences, what drives the encoding and memory differences shown in Experiments 1 and 2? There are several reasons why one would not see an interaction between sequence of study and type of feature on looking time.

First, most participants learned the categories, which requires participants in both sequencing conditions to attend to discriminating features between the items to a considerable degree. Second, looking at discriminating and characteristic features might be the result of a process of sequential comparison by which during interleaved study learners look preferentially to features that are different between successive items (mostly discriminating features; see Table 1), whereas for blocked study there would be no such preference (attending to all features, at least for the first category; see results of the feature prediction task of Experiment 1). As an example, if the Item 1 from category A in Table 1 is immediately followed by Item 4 of category B, there was a change from feature "2" to feature "1" in Dimension 1. Even though feature "1" is not predictive of categorization, according to SAT, it should be strongly encoded because it is a difference between different categories. Similarly, if Item 1 in category A is followed by Item 2 in category A, feature "2" in Dimension 1 should be strongly encoded because it is a similarity among items of the same category, whereas if Item 1 is followed by Item 4, feature "1" in Dimension 1 should be relatively ignored because it is not a similarity among items of the same category. Thus, differences in attention toward discriminative and characteristic features would result in similar preferences toward discriminative features that results not from a preference toward attending to discriminating features *per se*, but rather from the same process of information tracking in the context of different sequential statistical properties.

To investigate this possibility, we identified, for each subject and trial, the features that varied relative to the previous trial (different from previous trial), and those that were the same as in the previous trial (similar to previous trial) and calculated total time looking at each type of feature during the study phase. This score represents raw summed looking time. Importantly, sequential similarity/dissimilarity does not guarantee the discriminative value of the feature (e.g., there are differences between items of the same category as well as noncategory relevant differences between items of different categories; see Table 1). The results are depicted in the right panel of Figure 9.

Similarly to what was found with the previous analysis, learners spend overall more time looking at the items during interleaved study (M = 3,023, SD = 1,857) than blocked study (M = 2,099, SD = 991), $F(1, 66) = 37.13$, $p < .0001$, $\eta^2_G = 0.097$. Moreover, learners spend more time looking at properties that are different from those presented with the previous item (M = 2,928, SD = 1,720), than to features that are the same (M = 2,193, SD = 1,277), $F(1, 66) = 132.52$, $p < .0001$, $\eta^2_G = 0.064$. Importantly, there is a significant interaction between the sequence of study and type of feature, $F(1, 66) = 83.21$, $p < .0001$, $\eta^2_G = 0.040$. Pairwise comparisons using Holm correction indicate that learners look longer at different features than similar features during interleaved study, $t(66) = 12.14$, $p < .0001$, $d = 1.48$, but not during blocked study, $t(66) = 2.36$, $p = .062$, $d = 0.29$. Moreover, learners' total time looking at different features during interleaved study is significantly higher than for either different, $t(66) = 8.30$, $p < .0001$, $d = 1.01$, or similar, $t(66) = 9{,}27$, $p < .0001$, $d = 1.13$, features during blocked study. All other comparisons did not reach statistical significance, all ps $> .062$.

In sum, the results of this experiment indicate that (a) when correcting for baseline frequency differences, participants in this task spend overall more time looking at discriminative features than characteristic features, (b) participants spend overall less time looking at the features during blocked study than interleaved study, although this effect seems to be driven by the strong preference to attend to features that differ from those of the previously seen item during interleaved study, and (c) interleaved study results in a preference to attend to features that differ from the previous item studied, which will generally belong to a different category.

These results are consistent with the results of Experiment 1, Experiment 2, and SAT. During blocked study, learners do not show a bias toward differences between the current and previous item—this is likely to lead to strong encoding and memory of the characteristic features of the items, because these are more frequent, more often repeated among successive items and salient. During interleaved study, on the other hand, learners show a strong bias toward attending to what changed relative to the previous item studied. This is likely to lead to strong encoding of the discriminative features of the categories, because these are most often the differences between successive items of the different categories. This dynamic, we propose, is at the core of documented differences between the two sequences of study, and constitutes a mechanistic description of how learning takes place over time.

Finally, this process is probabilistic and other factors are likely to be at play. For example, in addition to sometimes being repeated across items of the same category, the discriminating features might have been more attended to and encoded during blocked study because their low frequency made them salient or surprising. Similarly, characteristic features constituted differences between items of different categories, even though they did not discriminate between categories, contributing to the encoding of characteristic features during interleaved study.

General Discussion

Overall, the results presented here show that different sequences of study result in different attentional patterns, encoding of different properties, and differential memory for the properties of the items. We believe these factors are intrinsically related. Through the use of different sequences of study, different attentional patterns are established that lead to different encoding and ultimately the creation of different memory traces.

When learners studied the items blocked by categories we saw an impact of the characteristic features of the categories studied on generalizations to novel items. These features were frequently associated with the category, but did not discriminate between categories. Conversely, when learners studied the categories interleaved changes in the same characteristic features had no impact

on novel item categorization. This suggests differences in how information was encoded in different sequences. Blocked study led to a relative greater emphasis on the characteristic properties of the category, whereas interleaved study emphasized discriminative features of the stimuli, ignoring nondiscriminative features, even if highly associated with the category. These results are consistent with previous research showing the independent roles of discriminative and characteristic features for categorization (Murphy & Ross, 2005), and that different types of tasks can affect which type of feature is encoded (Hoffman & Rehder, 2010; Markman & Ross, 2003; Yamauchi & Markman, 2000). In line with what has been shown for inference tasks, in the present studies blocked study, relative to interleaved study, led to the acquisition of features that have high category validity—the characteristic features of the objects. Conversely, interleaved study, relative to blocked study, led to the acquisition of features that have high cue validity—the discriminative features. Moreover, these results are consistent with previous research showing that different sequences of study lead to improved performance in different types of tasks (Carvalho & Goldstone, 2015a, 2017; Rawson et al., 2015); blocked study for inference-type tasks in which learners studied the item along with its category assignment and interleaved study for classification-type tasks during which learners had to "guess" the correct category assignment and were then given corrective feedback.

It is also interesting to note how powerful an effect characteristic features, even with virtually no cue validity, can have on categorization. Following blocked study, transfer to new items that lacked these characteristic features was negatively affected, even though the only features that predicted category assignment were still available. This is contrary to previous proposals and some common-sense interpretations of learning suggesting that category learning is about finding differences between categories. To the extent that learners use a similarity-based approach to novel item categorization (Kruschke, 1992; Love, Medin, & Gureckis, 2004; Nosofsky, 1986), all encoded features will influence categorization performance—even those with low predictive value for category assignment (Murphy & Ross, 2005; Rosch & Mervis, 1975; Wisniewski, 1995).

Differences in Encoding Are Related to Differences in Attention to Different Features

To further illustrate how changes to characteristic, nondiscriminative, features can have an impact on categorization accuracy and, critically, how this is connected to learners' attention to these features, we ran a simple simulation using the Generalized Context Model (GCM; Nosofsky, 1986). In GCM, stimuli are represented by points in a multidimensional space and a stimulus is classified as belonging to a category based on the similarity between that stimulus and all of the previously studied exemplars of that category relative to its similarity to all other categories' exemplars. Depending on the context (e.g., which other categories are being studied, their properties or the sequence of study), a specific dimension can become more or less relevant for categorization and therefore for how similar two stimuli are judged to be. In essence, in GCM selective attention parameters serve to stretch or shrink the dimensional space according to the learning context. Formally, the probability of categorizing an item i as belonging to a given category J is given by the summed similarity of that item to all the j exemplars of category J, divided by the summed similarity (S) of i to all the k exemplars of all the categories, K:

$$P(J \mid i) = \frac{\sum_{j=1}^{n} S_{ij}}{\sum_{k=1}^{n} S_{ik}}$$

To determine how similar two stimuli are, the model uses an exponential decaying function of distance. The similarity between items i and j is given by:

$$S_{ij} = e^{-cd_{ij}}$$

where d_{ij} is the attention-weighted distance between the two items. This calculation includes a freely estimated sensitivity parameter, c, that defines the rate by which similarity decays with distance, that is, the gradient of the similarity function. Smaller values of c correspond to an impact of more items to determine the similarity.

The distance between two stimuli in multidimensional space is a function of the differences between the two stimuli for all the dimensions considered. These differences are weighted by attention parameters ($w > 0$) that characterize how salient or relevant is each dimension. Thus, the distance between stimuli i and j is computed by:

$$d_{ij} = \left[\sum_{m=1}^{M} \left[w_m |x_{im} - x_{jm}|^r\right]\right]^{1/r}$$

where w_m is the attention allocated to dimension m, M is the total number of dimensions, and x_{im} and x_{jm} are the feature values of stimuli i and j on dimension m, respectively. A scaling parameter r is used to define the form of the distance metric. When $r = 1$, a city-block metric is used and when $r = 2$, a euclidean distance metric is used. Whereas w_m values are often free parameters fit to subject data, r is often defined by the type of stimuli used.

In our simple simulation, we had two categories, each with only one item: Category A with item x (1,1,1,1,1,1,1,1,3,3) and Category B with item y (1,1,1,1,1,1,1,2,4,4). Items in these two categories share most of their features. The first seven values correspond to characteristic (they are very common in the category), but nondiscriminative (items in both categories have them) features, whereas the last three numbers correspond to the values of discriminative features. We probe categorization of characteristic-preserved item x_p (3,1,1,1,1,1,1,1,5,5) and characteristic-changed item x_c (3,3,3,3,3,3,3,1,5,5) and we defined $r = 2$ and $c = 3$. Our proposal, consistent with the results presented in the three experiments in this paper, is that attention toward characteristic features is greater when stimuli are studied blocked than interleaved. To simulate this, we defined the weight for characteristic features $w_m = 0.5$ for blocked study and $w_m = 0.1$ for interleaved study. We defined $w_m = 1$ for the discriminative features. The results of the simulation is consistent with the empirical results presented in Experiments 1–3, and SAT predictions (see Figure 10).

As it can be seen from the graph in Figure 10, when more attention is paid to the characteristic features, as we propose is the case during blocked study, categorization of the characteristic-changed item is substantially lower relative to characteristic-preserved item. These results are robust to changes in the c parameter. This pattern is less marked when less attention is paid to the characteristic features, as we hypothesize is the case during

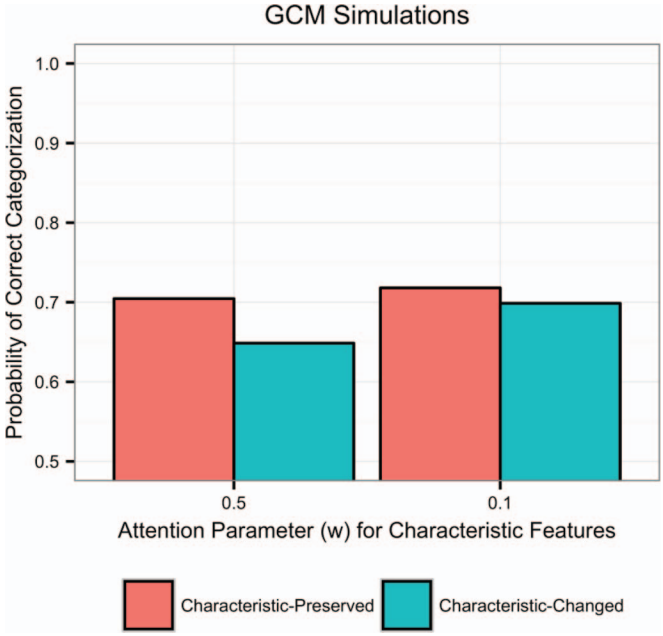

Figure 10. Results of model simulations using a simplified version of the generalized context model (GCM) categorization of two novel items: characteristic-preserved and characteristic-changed items. The attention parameter values used simulate the prediction that blocked study results in relatively higher attention toward characteristic features ($w = 0.5$) than interleaved study ($w = 0.1$). See the online article for the color version of this figure.

interleaved study. The resulting interaction in Figure 10 is similar to the one that was found in Experiment 1 and is shown in Figure 4. This simulation illustrates, using a successful and widely used model of categorization, how changes in the relevance of characteristic features for categorization are connected to changes in attention to and encoding of these features.

Differences in Encoding Affect What Information Is Available After Study

In addition to demonstrating that different attentional patterns lead to different information being encoded (and therefore relevant for categorization of novel items), the current work also takes a step further. With the inclusion of a memory task (Experiment 2) and eye tracking of where participants are looking while learning the categories (Experiment 3), we were able to link these differences in encoding to differences in what information learners have available following study in different sequences, and what in-the-moment sampling behaviors might lead to those differences. The features that learners attend to are shaped by the sequential statistics of the task—differences across trials during interleaved study and similarities across trials during blocked study. These differences in what is attended will, in turn, lead to differences in what is encoded. These ideas are at the center of SAT's mechanism for how sequential effects can shape category learning. It is also consistent with previous proposals of how category learning takes place across time (Jones & Sieck, 2003; Palmeri & Mack, 2015; Stewart, Brown, & Chater, 2002) and is context- and task-specific (Mack & Palmeri, 2015; Markman & Ross, 2003; Palmeri & Mack, 2015; Ross, 2000). Furthermore, this work and SAT are also congruent with recent neurophysiological evidence suggesting the important role of pattern completion for learning and the role of the hippocampus in not only providing details about past events but also about the relationship between events to create learning (Mack & Preston, 2016; Schlichting & Preston, 2015; Zeithamova, Schlichting, & Preston, 2012).

As briefly mentioned in the introduction, SAT proposes that category learning takes place as a series of pairwise contrasts between the stimulus currently being studied and the previously studied one, as well as their category assignments. If two successive stimuli belong to the same category, learners will tend to attend to similarities between the two, whereas if the stimuli are from different categories, learners will tend to attend to differences between the two. This process is consistent with previous evidence that learners show a recency bias during category learning (Jones, Love, & Maddox, 2006; Jones & Sieck, 2003; Stewart & Brown, 2004; Stewart et al., 2002; Zotov et al., 2011). Our manipulations and the current formulation of SAT pertain to only the immediately preceding item, but it is plausible that earlier items also have an influence (see, e.g., Stewart & Brown, 2004), presumably as a decreasing function of their temporal distance to the current item being studied.

Importantly, these differences or similarities between successively presented items are not necessarily globally relevant for categorization. Two items of different categories can show differences that are not discriminative of the categories, and similarities within categories might not discriminate between the categories. We believe this process to be at the core of differences found between interleaved and blocked sequences for successful category learning. During blocked study, learners are biased toward similarities shared among successive items of the same category. These similarities might prove useful in a category discrimination test if these characteristic features are also

highly discriminative of category membership (Carvalho & Goldstone, 2014b; Zulkiply & Burt, 2013), or when the test task does not emphasize discrimination between categories but a description of the most common features of the category (Carvalho & Albuquerque, 2012; Carvalho & Goldstone, 2017; de Zilva & Mitchell, 2012). However, blocked study can also lead learners astray when these temporally local similarities do not discriminate between categories, as in the present studies (see also Birnbaum et al., 2013; Rohrer & Taylor, 2007; Sana et al., 2017). Thus, SAT presents an account of how learning takes place across time and is shaped by sequence, being able to not only account for the benefits of either of the two sequences but also differences in what is attended to and encoded in different sequences (for a detailed discussion, see Carvalho & Goldstone, 2015b).

It is important to note that it is not our objective to argue that following sequential statistics (as SAT suggests) is the only driver of attentional behavior during learning. A general account of category learning must be more nuanced. Learners' attentional behavior is also guided, for example, by task-relevance (Rehder & Hoffman, 2005a) and novelty (Wang & Mitchell, 2011). Indeed, the fact that all learners attended to the discriminative features to a large degree argues that task relevance (for interleaved study) and/or relative novelty of discriminative features relative to characteristic ones (for blocked study), might have played a role in guiding learners' attention. However, our main proposal is that, in addition to these well-known attentional processes in categorization, sequential comparison of features will also have a powerful effect on attention. Moreover, the results of Experiment 2 and Experiment 3 also suggest that the very nature of the task might influence overall attentional patterns, with decreased overall memory and attention following blocked compared to interleaved study.

Concluding Remarks

The results presented here have theoretical and practical implications. At a theoretical level, these results, and our theoretical interpretation, contribute to the ever-growing interest of the situated nature of category learning and its online processes (e.g., Carvalho & Goldstone, 2014b; Palmeri & Mack, 2015; Qian & Aslin, 2014). Moreover, our theory can provide a parsimonious conceptualization of not only the results presented here, but also how the different effects of different sequences result from a common learning process. It might be tempting to think that there are two different learning processes, one for interleaved study and another for blocked study (see, e.g., Birnbaum et al., 2013; Rohrer, Dedrick, & Stershic, 2015; Sana et al., 2017) or to define different attentional weights *a posteriori* for different sequences. However, SAT is a unified account that does not require one to postulate separate learning processes for each sequence, but a single one that naturally gives rise to differences between sequences. At a practical level, these results reiterate the important role that a carefully organized sequence of study can play for learning outcomes (e.g., Kellman, 2013; Koedinger, Booth, & Klahr, 2013; Li et al., 2013; Mettler & Kellman, 2013; Rau, Aleven, & Rummel, 2013). Moreover, these results emphasize the importance of considering the whole learning and testing situation when choosing how to sequence one's learning. An understanding of the learning situation as a whole and how a unified learning mechanism is changed by it allows for recommendations beyond crude one-size-fits-all strategies, giving instructors and learners flexibility to improve learning across multiple contexts.

References

Anderson, J. R. (1991). The adaptive nature of human categorization. *Psychological Review, 98,* 409–429. http://dx.doi.org/10.1037/0033-295X.98.3.409

Baddeley, A. D., & Hitch, G. J. (1977). Recency re-examined. *Attention and Performance, 6,* 647–667.

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

Blair, M. R., Watson, M. R., & Meier, K. M. (2009). Errors, efficiency, and the interplay between attention and category learning. *Cognition, 112,* 330–336. http://dx.doi.org/10.1016/j.cognition.2009.04.008

Blair, M. R., Watson, M. R., Walshe, R. C., & Maj, F. (2009). Extremely selective attention: Eye-tracking studies of the dynamic allocation of attention to stimulus features in categorization. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 35,* 1196–1206. http://dx.doi.org/10.1037/a0016272

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,* 245–248. http://dx.doi.org/10.1080/00220671.1981.10885317

Brady, F. (2008). The contextual interference effect and sport skills. *Perceptual and Motor Skills, 106,* 461–472. http://dx.doi.org/10.2466/pms.106.2.461-472

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

Carvalho, P. F., & Albuquerque, P. B. (2012). Memory encoding of stimulus features in human perceptual learning. *Journal of Cognitive Psychology, 24,* 654–664. http://dx.doi.org/10.1080/20445911.2012.675322

Carvalho, P. F., & Goldstone, R. L. (2011). Sequential similarity and comparison effects in category learning. In L. Carlson, C. Holscher, & T. Shipley (Eds.), *Proceedings of the 33rd conference of the Cognitive Science Society* (pp. 2977–2982). Austin, TX: Cognitive Science Society.

Carvalho, P. F., & Goldstone, R. L. (2014a). Effects of interleaved and blocked study on delayed test of category learning generalization. *Frontiers in Psychology, 5,* 936. http://dx.doi.org/10.3389/fpsyg.2014.00936

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

Carvalho, P. F., & Goldstone, R. L. (2015a). The benefits of interleaved and blocked study: Different tasks benefit from different schedules of study. *Psychonomic Bulletin & Review, 22,* 281–288. http://dx.doi.org/10.3758/s13423-014-0676-4

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

Carvalho, P. F., & Goldstone, R. L. (2017). *The most efficient sequence of study depends on the type of test*. Manuscript in preparation.

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,* 354–380. http://dx.doi.org/10.1037/0033-2909.132.3.354

Chen, L., Meier, K. M., Blair, M. R., Watson, M. R., & Wood, M. J. (2013). Temporal characteristics of overt attentional behavior during

category learning. *Attention, Perception, & Psychophysics, 75,* 244–256. http://dx.doi.org/10.3758/s13414-012-0395-8

Clapper, J. P. (2014). The impact of training sequence and between-category similarity on unsupervised induction. *The Quarterly Journal of Experimental Psychology, 68,* 1–55.

Corcoran, K., Epstude, K., Damisch, L., & Mussweiler, T. (2011). Fast similarities: Efficiency advantages of similarity-focused comparisons. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 37,* 1280–1286. http://dx.doi.org/10.1037/a0023922

Crowder, R. G. (1976). *Principles of learning and memory*. Hillsdale, NJ: Erlbaum.

de Zilva, D., & Mitchell, C. J. (2012). Effects of exposure on discrimination of similar stimuli and on memory for their unique and common features. *The Quarterly Journal of Experimental Psychology: Human Experimental Psychology, 65,* 1123–1138. http://dx.doi.org/10.1080/17470218.2011.644304

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,* 795–805. http://dx.doi.org/10.1037/0021-9010.84.5.795

Elio, R., & Anderson, J. R. (1984). The effects of information order and learning mode on schema abstraction. *Memory & Cognition, 12,* 20–30. http://dx.doi.org/10.3758/BF03196994

Glanzer, M., Adams, J. K., Iverson, G. J., & Kim, K. (1993). The regularities of recognition memory. *Psychological Review, 100,* 546–567. http://dx.doi.org/10.1037/0033-295X.100.3.546

Glanzer, M., & Cunitz, A. R. (1966). Two storage mechanisms in free recall. *Journal of Verbal Learning & Verbal Behavior, 5,* 351–360. http://dx.doi.org/10.1016/S0022-5371(66)80044-0

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

Helsdingen, A. S., van Gog, T., & van Merriënboer, J. J. G. (2011). The effects of practice schedule on learning a complex judgment task. *Learning and Instruction, 21,* 126–136. http://dx.doi.org/10.1016/j.learninstruc.2009.12.001

Hoffman, A. B., & Rehder, B. (2010). The costs of supervised classification: The effect of learning task on conceptual flexibility. *Journal of Experimental Psychology: General, 139,* 319–340. http://dx.doi.org/10.1037/a0019042

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics: Theory and Applications, 6,* 65–70.

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,* 138–149. http://dx.doi.org/10.1086/374692

Jones, M., Love, B. C., & Maddox, W. T. (2006). Recency effects as a window to generalization: Separating decisional and perceptual sequential effects in category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 32,* 316–332. http://dx.doi.org/10.1037/0278-7393.32.3.316

Jones, M., & Sieck, W. R. (2003). Learning myopia: An adaptive recency effect in category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 29,* 626–640. http://dx.doi.org/10.1037/0278-7393.29.4.626

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

Kellman, P. J. (2013). Adaptive and perceptual learning technologies in medical education and training. *Military Medicine, 178*(10, Suppl), 98–106. http://dx.doi.org/10.7205/MILMED-D-13-00218

Koedinger, K. R., Booth, J. L., & Klahr, D. (2013). Education research. Instructional complexity and the science to constrain it. *Science, 342,* 935–937. http://dx.doi.org/10.1126/science.1238056

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

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

Kost, A. S., Carvalho, P. F., & Goldstone, R. L. (2015). Can you repeat that? The effect of item repetition on interleaved and blocked study. In D. C. Noelle, R. Dale, A. S. Warlaumont, J. Yoshimi, T. Matlock, C. D. Jennings, & P. P. Maglio (Ed.), *Proceedings of the 37th annual meeting of the Cognitive Science Society* (pp. 1189–1194). Austin, TX: Cognitive Science Society.

Kruschke, J. K. (1992). ALCOVE: An exemplar-based connectionist model of category learning. *Psychological Review, 99,* 22–44. http://dx.doi.org/10.1037/0033-295X.99.1.22

Kruschke, J. K., Kappenman, E. S., & Hetrick, W. P. (2005). Eye gaze and individual differences consistent with learned attention in associative blocking and highlighting. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 31,* 830–845. http://dx.doi.org/10.1037/0278-7393.31.5.830

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

Li, N., Cohen, W. W., & Koedinger, K. R. (2013). Problem order implications for learning. *International Journal of Artificial Intelligence in Education, 23,* 71–93. http://dx.doi.org/10.1007/s40593-013-0005-5

Lipsitt, L. P. (1961). Simultaneous and successive discrimination learning in children. *Child Development, 32,* 337–347.

Love, B. C., Medin, D. L., & Gureckis, T. M. (2004). SUSTAIN: A network model of category learning. *Psychological Review, 111,* 309–332. http://dx.doi.org/10.1037/0033-295X.111.2.309

Mack, M. L., & Palmeri, T. J. (2015). The dynamics of categorization: Unraveling rapid categorization. *Journal of Experimental Psychology: General, 144,* 551–569. http://dx.doi.org/10.1037/a0039184

Mack, M. L., & Preston, A. R. (2016). Decisions about the past are guided by reinstatement of specific memories in the hippocampus and perirhinal cortex. *NeuroImage, 127,* 144–157. http://dx.doi.org/10.1016/j.neuroimage.2015.12.015

Malmberg, K. J., & Xu, J. (2006). The influence of averaging and noisy decision strategies on the recognition memory ROC. *Psychonomic Bulletin & Review, 13,* 99–105. http://dx.doi.org/10.3758/BF03193819

Markman, A. B., & Ross, B. H. (2003). Category use and category learning. *Psychological Bulletin, 129,* 592–613. http://dx.doi.org/10.1037/0033-2909.129.4.592

McDaniel, M. A., Fadler, C. L., & Pashler, H. (2013). Effects of spaced versus massed training in function learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 39,* 1417–1432. http://dx.doi.org/10.1037/a0032184

Medin, D. L. (1983). Structural principles of categorization. In T. J. Tighe & B. E. Shepp (Eds.), *Perception, cognition, and development: Interactional analyses* (pp. 203–230). Hillsdale, NJ: Erlbaum.

Mettler, E., & Kellman, P. J. (2013). Adaptive response-time-based category sequencing in perceptual learning. *Vision Research, 99,* 111–123. http://dx.doi.org/10.1016/j.visres.2013.12.009

Morey, R. D., Pratte, M. S., & Rouder, J. N. (2008). Problematic effects of aggregation in z ROC analysis and a hierarchical modeling solution. *Journal of Mathematical Psychology, 52,* 376–388. http://dx.doi.org/10.1016/j.jmp.2008.02.001

Murphy, G. L. (1982). Cue validity and levels of categorization. *Psychological Bulletin, 91,* 174–177. http://dx.doi.org/10.1037/0033-2909.91.1.174

Murphy, G. L., & Ross, B. H. (2005). The two faces of typicality in category-based induction. *Cognition, 95,* 175–200. http://dx.doi.org/10.1016/j.cognition.2004.01.009

Murray, J. T. (1983). *Spacing phenomena in human memory: A study-phase retrieval interpretation*. Los Angeles, CA: University of California.

Nosofsky, R. M. (1986). Attention, similarity, and the identification-categorization relationship. *Journal of Experimental Psychology: General, 115,* 39–57. http://dx.doi.org/10.1037/0096-3445.115.1.39

Palmeri, T. J., & Mack, M. L. (2015). How experimental trial context affects perceptual categorization. *Frontiers in Psychology, 6,* 180.

Phillips, W. A., & Christie, D. F. M. (1977). Components of visual memory. *The Quarterly Journal of Experimental Psychology, 29,* 117–133. http://dx.doi.org/10.1080/00335557743000080

Qian, T., & Aslin, R. N. (2014). Learning bundles of stimuli renders stimulus order as a cue, not a confound. *Proceedings of the National Academy of Sciences of the United States of America, 111,* 14400–14405. http://dx.doi.org/10.1073/pnas.1416109111

Ratcliff, R., Sheu, C.-F., & Gronlund, S. D. (1992). Testing global memory models using ROC curves. *Psychological Review, 99,* 518–535. http://dx.doi.org/10.1037/0033-295X.99.3.518

Rau, M. A., Aleven, V., & Rummel, N. (2013). Interleaved practice in multi-dimensional learning tasks: Which dimension should we interleave? *Learning and Instruction, 23,* 98–114. http://dx.doi.org/10.1016/j.learninstruc.2012.07.003

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,* 483–504. http://dx.doi.org/10.1007/s10648-014-9273-3

R Core Team. (2013). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from http://www.r-project.org/

Rehder, B., & Hoffman, A. B. (2005a). Eyetracking and selective attention in category learning. *Cognitive Psychology, 51,* 1–41. http://dx.doi.org/10.1016/j.cogpsych.2004.11.001

Rehder, B., & Hoffman, A. B. (2005b). Thirty-something categorization results explained: Selective attention, eyetracking, and models of category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 31,* 811–829. http://dx.doi.org/10.1037/0278-7393.31.5.811

Rohrer, D. (2009). The effects of spacing and mixing practice problems. *Journal for Research in Mathematics Education, 40,* 14.

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

Rohrer, D., Dedrick, R. F., & Stershic, S. (2015). Interleaved practice improves mathematics learning. *Journal of Educational Psychology, 107,* 900–908.

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

Rosch, E., & Mervis, C. B. (1975). Family resemblances: Studies in the internal structure of categories. *Cognitive Psychology, 7,* 573–605. http://dx.doi.org/10.1016/0010-0285(75)90024-9

Ross, B. H. (2000). The effects of category use on learned categories. *Memory & Cognition, 28,* 51–63. http://dx.doi.org/10.3758/BF03211576

Samuels, S. J. (1969). Effect of Simultaneous versus successive discrimination training on paired-associate learning. *Journal of Educational Psychology, 60,* 46–48. http://dx.doi.org/10.1037/h0026671

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

Sandhofer, C. M., & Doumas, L. A. A. (2008). Order of Presentation Effects in Learning Color Categories. *Journal of Cognition and Development, 9,* 194–221. http://dx.doi.org/10.1080/15248370802022639

Schlichting, M. L., & Preston, A. R. (2015). Memory integration: Neural mechanisms and implications for behavior. *Current Opinion in Behavioral Sciences, 1,* 1–8. http://dx.doi.org/10.1016/j.cobeha.2014.07.005

Stanislaw, H., & Todorov, N. (1999). Calculation of signal detection theory measures. *Behavior Research Methods, Instruments & Computers, 31,* 137–149. http://dx.doi.org/10.3758/BF03207704

Ste-Marie, D. M., Clark, S. E., Findlay, L. C., & Latimer, A. E. (2004). High levels of contextual interference enhance handwriting skill acquisition. *Journal of Motor Behavior, 36,* 115–126. http://dx.doi.org/10.3200/JMBR.36.1.115-126

Stewart, N., & Brown, G. D. A. (2004). Sequence effects in the categorization of tones varying in frequency. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 30,* 416–430. http://dx.doi.org/10.1037/0278-7393.30.2.416

Stewart, N., Brown, G. D. A., & Chater, N. (2002). Sequence effects in categorization of simple perceptual stimuli. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 28,* 3–11. http://dx.doi.org/10.1037/0278-7393.28.1.3

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

Wang, T., & Mitchell, C. J. (2011). Attention and relative novelty in human perceptual learning. *Journal of Experimental Psychology: Animal Behavior Processes, 37,* 436–445. http://dx.doi.org/10.1037/a0023104

Wisniewski, E. J. (1995). Prior knowledge and functionally relevant features in concept learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 21,* 449–468. http://dx.doi.org/10.1037/0278-7393.21.2.449

Yamauchi, T., & Markman, A. B. (2000). Inference using categories. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 26,* 776–795. http://dx.doi.org/10.1037/0278-7393.26.3.776

Yonelinas, A. P., & Parks, C. M. (2007). Receiver operating characteristics (ROCs) in recognition memory: A review. *Psychological Bulletin, 133,* 800–832. http://dx.doi.org/10.1037/0033-2909.133.5.800

Zeithamova, D., & Maddox, W. T. (2009). Learning mode and exemplar sequencing in unsupervised category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 35,* 731–741. http://dx.doi.org/10.1037/a0015005

Zeithamova, D., Schlichting, M. L., & Preston, A. R. (2012). The hippocampus and inferential reasoning: Building memories to navigate future decisions. *Frontiers in Human Neuroscience, 6,* 70.

Zotov, V., Jones, M. N., & Mewhort, D. J. K. (2011). Contrast and assimilation in categorization and exemplar production. *Attention, Perception, & Psychophysics, 73,* 621–639. http://dx.doi.org/10.3758/s13414-010-0036-z

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

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

(*Appendices follow*)

Appendix A

Averaged Receiver Operant Characteristics (ROC) Curves for the Memory Task of Experiment 2

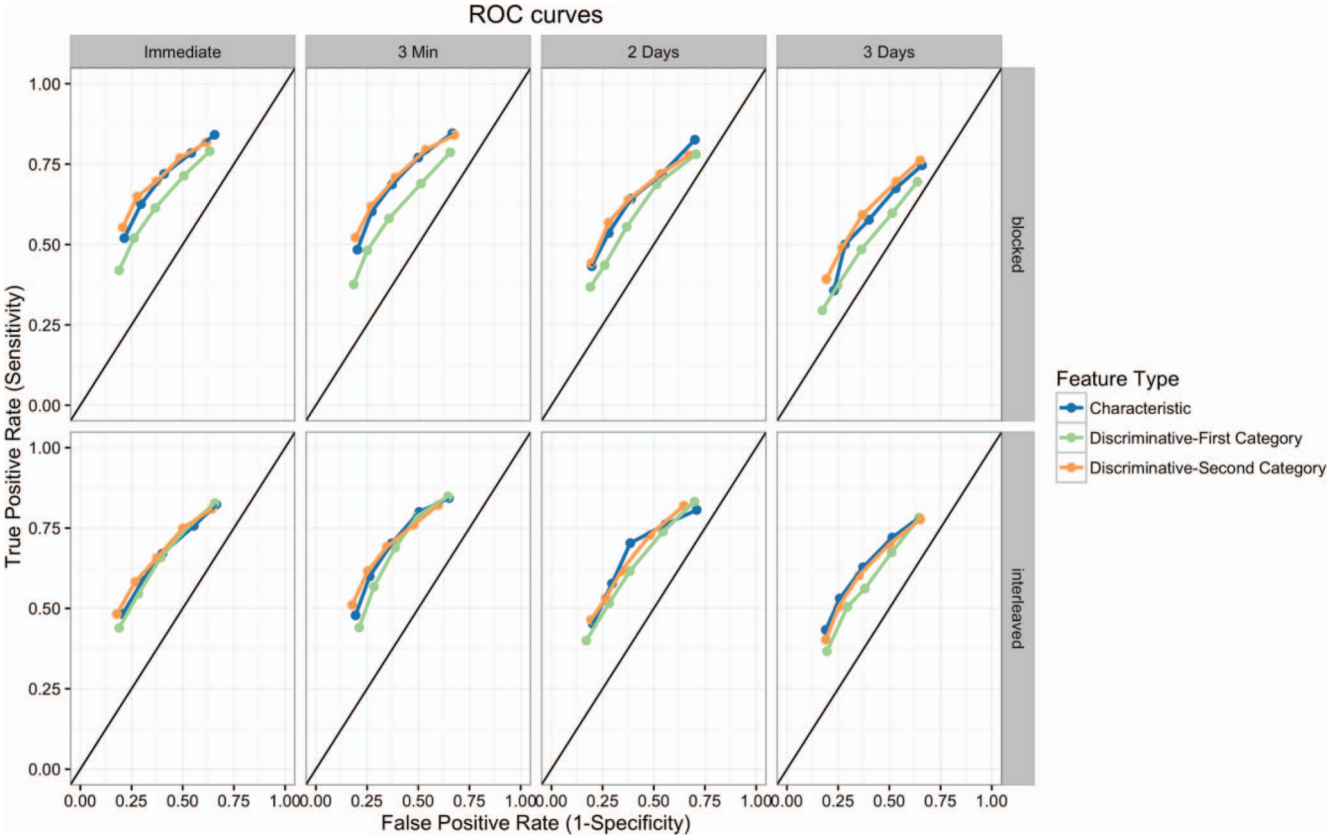

Figure A1. Receiver operant characteristics (ROCs) for the three types of features across both study sequences and retention intervals plotted in probability space. Points represent average hit and false-alarm rates across participants. The diagonal represents chance responding. See the online article for the color version of this figure.

(*Appendices continue*)

Appendix B

Averaged Receiver Operant Characteristics (ROC) curves Plotted in z-space for the Memory Task of Experiment 2

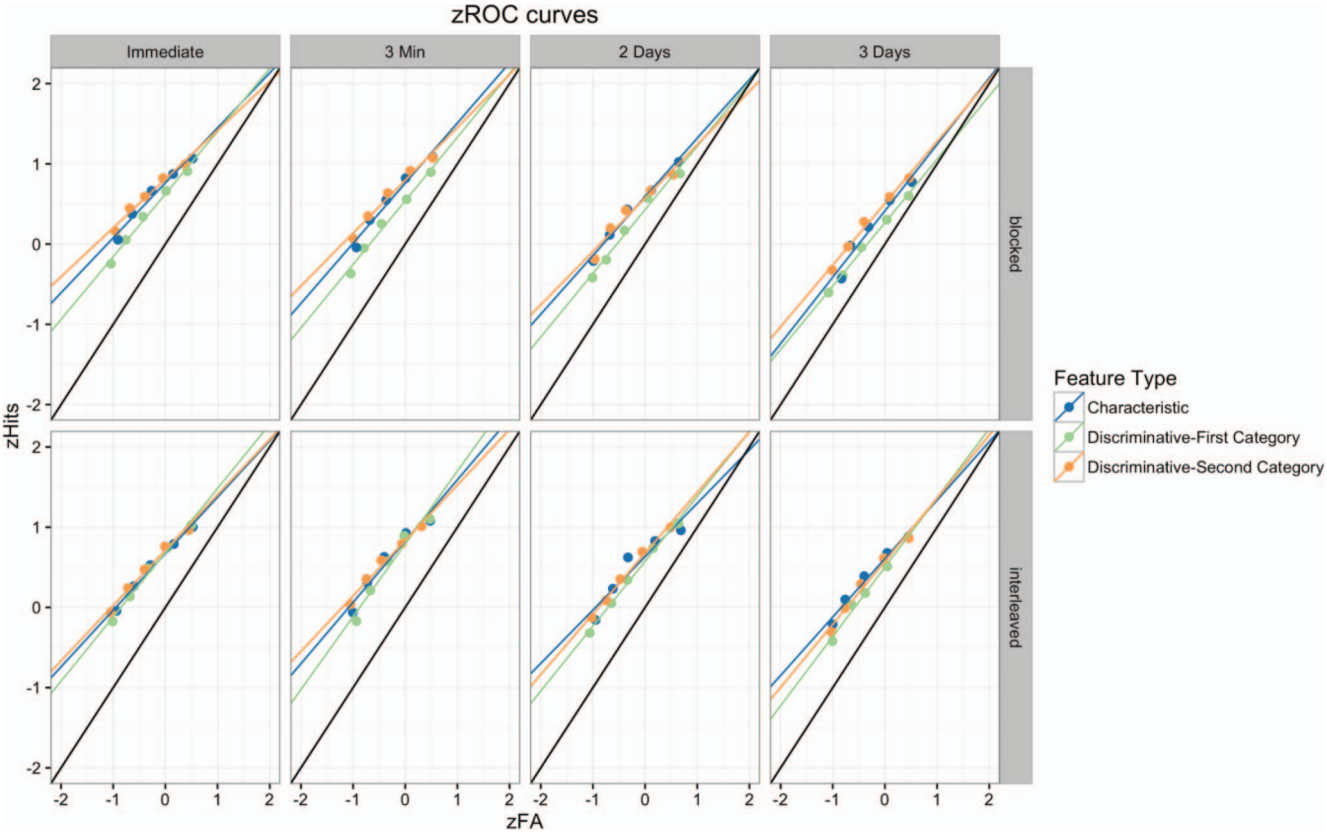

Figure A2. Receiver operant characteristics plotted in z-space (zROCs) for the different types of features across both study sequences and retention intervals. Points represent average zHit and zFA rates across participants. Lines represent the best-fitting lines for each group of points using conventional regression methods. See the online article for the color version of this figure.

Received July 5, 2016
Revision received December 13, 2016
Accepted February 9, 2017 ■