# A computational model of context-dependent encodings during category learning

Paulo F. Carvalho

Carnegie Mellon University

Robert L. Goldstone

Indiana University

## Author Note

Paulo F. Carvalho, Human-Computer Interaction Institute, Carnegie Mellon University, Pittsburgh, PA

Robert L. Goldstone. Department of Psychological and Brain Sciences and Cognitive Science Program, Indiana University, Bloomington, IN.

Correspondence concerning this article should be addressed to Paulo Carvalho, Human-Computer Interaction Institute, Carnegie Mellon University, 5000 Forbes Ave, Pittsburgh, PA 15213, USA. Email: pcarvalh@andrew.cmu.edu

This work is part of a doctoral dissertation submitted by the first author to Indiana University as partial fulfilment of the degree of Doctor of Philosophy. Parts of this work were supported by the National Science Foundation [grant #1824257 to PC], the Department of Education [IES grant # R305A1100060 to RG], and the Portuguese Foundation for Science and Technology, co-sponsored by the European social fund [Graduate Training Fellowship grant SFRH/BD/78083/2011 to PC].

## Abstract

Although current exemplar models of category learning are flexible and can capture how different features are emphasized for different categories, they still lack in the flexibility to adapt to local changes in category learning, such as the effect of different sequences of study. In this paper we introduce a new model of category learning, the Sequential Attention Theory Model (SAT-M), in which the encoding of each presented item is influenced not only by its category assignment (global context) as in other exemplar models, but also by how its properties relate to the properties of temporally neighboring items (local context). We demonstrate that SAT-M captures the effect of local context and predicts not only learning outcomes but also learners' attentional patterns during training.

*Keywords*: category learning models; sequencing; interleaving; attention; encoding

# A computational model of context-dependent encodings during category learning

Our categorization decisions are often based on a subset of potential features. For example, when at the supermarket trying to pick the perfectly ripe avocado for dinner, we look for a softer, darker avocado. Although avocados can differ on many properties, color seems to be particularly relevant for ripeness. Not all features are equally important, and humans are particularly good at discerning what matters from what does not. Moreover, discerning the features that matter is context dependent. For example, when looking for a book among bowls, shape is a critical feature.  However, when looking for a particular book among other books that is unlikely to be the case. How do we flexibly learn what to attend to and encode?

Extant theories and models of human and artificial category learning try to answer this question by taking the attentional flexibility of human categorization into account (e.g., Bareiss et al., 1990; Nosofsky, 1986). One common way to address the variable importance of different features is to assume that different properties of to-be-categorized items have different weights that modulate their overall impact on categorization decisions (Anderson et al., 1997; Ashby et al., 2011; Hintzman, 1984; Kruschke, 1992; Love et al., 2004; Nosofsky, 2011; Thiessen & Pavlik, 2013).

For example, the Generalized Context Model (GCM; Nosofsky, 1986) proposes a set of selective-attention weights that represent the learned strategy of attending more to some properties than others when making decisions about stimuli represented in a multidimensional space. In GCM the attentional weights are either free parameters fit to the data or assigned plausible values to approximate human behavior in a categorization task. Although the assumption is that these attentional weights to different dimensions are learned, GCM does not include an explicit process for how attention changes over time during learning.

Other models, however, have addressed this question. For example, the Attention Learning COVEring map (ALCOVE) model of categorization (Kruschke, 1992) includes a process through which attentional weights are learned during learning. On each learning trial, ALCOVE computes the discrepancy between the classification provided and the correct classification and adjusts the attention weights to reduce the classification error. Other models provide similar mechanisms for learned attentional weights, for example, SUSTAIN (Love et al., 2004) includes an attentional tuning algorithm whereby attentional weights are updated to provide greater salience to relevant feature dimensions and MINERVA 2 (Hintzman, 1984) includes an abstraction mechanism through which items are re-encoded emphasizing learning-relevant information (for a further specification of attentional processes in MINERVA see iMINERVA, Thiessen & Pavlik, 2013).

## Global and Local pressures on category learning

Although taking steps toward a mechanism by which we learn what to attend to and flexibly adapt to the properties of the task and the stimuli being learned, these models still fall short when considering learning as a process through time. Broadly speaking, even in models with a learning process such as ALCOVE and SUSTAIN, attention is updated to change the salience of dimensions based on their overall relevance for categorization, that is, taking into account the global context of all of items in the same category.

Current formal and informal models of categorization assign attentional weights during learning to facilitate human flexibility in categorization. However, there are two major issues with this approach. First, it implies that learners have access to the entire set of previously seen examples when making a local categorization decision and that all items and all their features are equally important, which at face value seems unlikely. Indeed, previous research has demonstrated a "recency effect" in categorization, whereby more recently studied items have a larger impact on categorization decisions than earlier ones (Jones et al., 2006; Jones & Sieck, 2003). One extrapolation of these findings is that, at the extremes, some previously seen items might have a small or negligible effect on categorization decisions.

Second, it implies that category learning is guided only by overall attention weights that might or might not be relevant for the task at hand. That is, when deciding which features to attend during learning, existing models optimize for overall categorization decisions: is feature *x* predictive of category *X*? Although feature *x* might, overall, predict category *X*, in the moment-to-moment context of the other items studied (as in the book/bowl example above), that might not be the case: a feature that generally predicts *X* might not be relevant in the local context of a series of items that do not have that feature. Although feature X might be relevant to categorization, if that is not the case at the local context level of the recent few items seen, then it is unlikely to receive much attention.

If instead we think of category learning as a process over time in which the learner is concerned only in solving the local task of making sense of how a current item and its temporal neighbors are grouped and in which every item has local contextual pressures on attention and encoding (from other stimuli, variation in task, etc.), we can create more flexible and informationally leaner models of learning. In this view, human learning as a process in time is the result of pressures at a small timescale. The *local context* of categorization is not only what category an item belongs to, but how that item fits with items recently categorized into one or another category. For example, a feature that predicts categorization overall (global context) might be missing from a specific exemplar or be present in an exemplar of the opposite category just presented in the previous trial (local context), in which case a different feature might be more relevant *locally*. This is a critical distinction in a world where categorization decisions are potentially seldomly based on all the information available about each category but instead only a few cases, potentially because of memory or expertise constraints (Elio & Anderson, 1984; Jones & Sieck, 2003). While models like EBRW (Nosofsky & Palmeri, 1997) selectively sample a restricted set of previously seen examples according to their similarity to an item to be categorized, we are proposing to selectively sample examples according to their temporal proximity to the item.

Consistent with this view, there is evidence that what information is attended to and encoded depends not only on its global context (what category it belongs to), which current models of categorization focus on, but also on its local context (neighbor items), which most current formal and informal models of categorization ignore. For example, Aha and Goldstone (1992) demonstrated that humans are able to learn categories that require attributes to be weighted differently for different category *exemplars*. That is, human categorization is sensitive to the local context of an exemplar relative to its stimulus space.

Another example of how human learning is sensitive to not only the global context in which category an item belongs to but also the local context of which other items are studied in proximity comes from a wealth of studies demonstrating sequencing effects in category learning. The sequence of events has been shown to have an impact on what we learn (Goldstone, 1996; Schyns & Rodet, 1997) and how well we learn it (Kornell & Bjork, 2008; Wahlheim et al., 2011). Because different sequences will result in the same item having different local neighbors for each learning event, this is both evidence that human categorization is sensitive to local variation and an ideal test situation for the adequacy of our model.

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).

To account for some of these findings, Carvalho and Goldstone (Carvalho & Goldstone, 2014b, 2015a, 2015b, 2017), have proposed the Sequential Attention Theory (SAT; Carvalho and Goldstone, 2017). According to SAT, one of the ways in which the sequence of learning influences learning is by creating pressures on what information is attended and encoded. Specifically, Carvalho and Goldstone, proposed – and empirically demonstrated (Carvalho & Goldstone, 2014b, 2014a, 2015a) – that studying items of the same category consecutively (blocked study) makes similarities between temporally neighboring items more salient and more likely to be used for categorization, whereas studying items of different categories consecutively (interleaved study) makes their differences more salient and more likely to be used for categorization. This proposal is able to capture many findings in the literature showing, for example, that interleaved study of categories improves learning of similar categories (for which detecting differences are particularly important), whereas blocked study of categories improves learning of dissimilar categories (for which detecting similarities within each category is important; for a metaanalysis see Brunmair & Richter, in press).

In sum, one of the critical pressures affecting where attention will be directed during category learning is likely to be the local context of the other items experienced in proximity (temporal, physical, etc.). Local influences go beyond the global context of whether a dimension or property has been relevant for categorization overall. The effect of both global and local influences on categorization is likely to be exerted through selective attending and encoding different properties with different experience, resulting in different properties being attended such as in the avocado example in the start of this paper. However, previous models implementing learning mechanisms that capture this phenomenon have focused only on the global context. In this paper we present a new model that considers the local context of categorization as critical, focusing on the local temporal context.

## A new model of category learning: SAT-M

We propose the Sequential Attention Theory Model (SAT-M), which computationally specifies and extends the SAT framework proposed by Carvalho and Goldstone (2017). SAT-M is a new exemplar model of human categorization based on GCM in which how items are encoded depends not only on global attentional weights that maximize correct categorization, but also on the immediate context during learning (i.e., the other stimuli studied in close proximity). If humans are, in fact, sensitive to local context in how they encode information, then SAT-M should provide a better characterization of learning behavior and category generalization, especially in situations where local context is particularly variable. One such situation is when one varies the sequence of exemplars during learning.

### Overview of SAT-M

In SAT-M, as in other exemplar models of categorization, each experienced item is represented by its value properties. During learning, each stimulus is stored along with its category assignment and encoding weights for each of its features. Later classification of test items, both trained and novel, depends on their similarity to the aggregate properties of the previously trained items, where similarity is based on how distant these test items are from the previously studied items in the categorization space. A test item will tend to be placed into a category to the extent that it is more similar to items in that category than other categories.

Unlike other exemplar models of categorization, however, in SAT-M, the representation of studied items is influenced not only by its properties and the relative relevance of each property for categorization, but also by how those properties and the category assignment vary from prior items seen before it. Thus, unlike other exemplar models, the stored representation is biased towards features that were relevant for its categorization in the context of the other item studied in temporal proximity. This is achieved by storing for each item feature an encoding weight that modulates the impact that that feature will have to determine the item's similarity to new items at test. For example, if two similar items of the same category are seen in close temporal proximity, their similarities will be more strongly encoded and these stimuli will be seen as more similar than if they had been studied further apart in time.

Our proposal is that by differentially encoding different properties of the stimuli from trial to trial, we are creating different attentional pressures and the history of categorization will result in different features playing a larger or smaller role during learning and future categorization. In the next section we discuss how the model works in more detail, its individual components, and principles.

**SAT-M principles and implementation**

As mentioned above, SAT-M builds on previous exemplar models. Specifically, SAT-M shares many of its principles and assumptions from the General Context Model (Nosofsky, 1986). As such, we will describe the general mechanisms of GCM and note how SAT-M deviates from it.

The principles of GCM and SAT-M can be divided into three main parts: item encoding, item similarity, and decision process. We will describe the model starting from its observable behavior, how it makes a categorization decision, and work backwards to the stimulus encoding that is used as a basis for the categorization decision. To foreshadow, categorization decisions are based on the perceived similarity of the to-be-categorized stimuli to all previously categorized stimuli and their respective category assignments. The similarity between stimuli is, in turn, inversely related to distance between their integrated distance along all of their component features. Finally, in SAT-M, unlikely GCM, the distance between stimuli is a function of each stimulus's properties as influenced by their temporal context (i.e., the properties of adjacent neighboring stimuli).

**Categorization decision.** To decide how to categorize a given item *y*, GCM takes into account how similar item *y* is to all other items encoded. Item *y* will tend to be categorized as belonging to the category containing the greatest number of items similar to *y*. The categorization probability uses a ratio rule, and is an application of Luce's choice rule (Luce, 1963; Shepard, 1965). The probability of categorizing an item *y* as belonging to a given category *A* is given by the summed similarity of that item to all the *x* exemplars of category *A*, divided by the summed similarity of *y* to all the *k* exemplars of all the categories, *K*:

$$\boldsymbol{p(A \mid y)} = \frac{\beta_A[\sum_x \boldsymbol{s_{xy}}]^\gamma}{\sum_K \beta_K[\sum_{\boldsymbol{k}} \boldsymbol{s_{ky}}]^\gamma} \qquad (1)$$

where $S_{xy}$ denotes the similarity between item *y* and item *x*. $\beta_A > 0$ represents the bias towards responding with category *A*. A response-scaling parameter, $\gamma$. When $\gamma=1$ the model responds more probabilistically by "probability matching", with greater values of $\gamma$, the model's behavior is more deterministic towards the category with greater summed similarity (Nosofsky & Zaki, 2002).

**Item similarity**. To determine how similar two items are, the model uses an exponentially decaying function of distance. The similarity between items *x* and *y* is given by:

$$\boldsymbol{s_{xy} = e^{-c\ d(x,y)^{\rho}}} \qquad (2)$$

where *d(x,y)* 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, i.e., the gradient of the similarity function. As *c* increases, categorization decisions are disproportionately influenced by trained items that are very close to the test item. As *c* decreases, then categorization is more equitably determined by all trained items, close or far, from the test item. Finally, the shape of the function relating distance and similarity is defined by a parameter $\rho$. When $\rho = 1$, there is an exponential relation between similarity and distance, while when $\rho = 2$, there is a Gaussian relation between the two. This parameter is often set in advance, and values greater than 1 are often used for highly confusable stimuli (Nosofsky, 2011).

**Distance in multidimensional space.** The absolute difference between two items is determined for each dimension in the multidimensional space. This difference is weighted by an attention parameter ($\boldsymbol{\omega_i} > 0$) that characterizes the global salience or relevance of dimension *i,* similar to previous exemplar models. The attention-weighted sum of the differences for all dimensions is the total distance between the two stimuli. Thus, the distance between stimuli *x* and *y* is computed by:

$$\boldsymbol{d}\,(\boldsymbol{x},\boldsymbol{y}) = \left[\sum_{\boldsymbol{i}} \boldsymbol{\omega_i}\, \boldsymbol{\varepsilon}_i^x\, |\boldsymbol{x_i} - \boldsymbol{y_i}|^{\boldsymbol{r}}\right]^{\frac{1}{r}} \qquad (3)$$

where $\boldsymbol{\omega_i}$ is the attention allocated to Dimension *i* and $x_i$ and $y_i$ are the feature values of stimuli *x* and *y* on dimension *i*. Notice that when $\boldsymbol{\omega}$ is high, that dimension will have a greater influence in the distance calculations, whereas smaller values of $\boldsymbol{\omega}$ will render any differences along that dimension less influential. This parameter can reflect learned information about which dimensions are relevant for categorization and which are not. A scaling parameter *r* is used to define the form of the distance metric. Whereas $\rho$ defines the relation between distance and similarity in calculating items' similarities (with $\rho$=1 implementing an exponential relation while $\rho$=2 implements a Gaussian relation), *r* defines the form of the distance metric when calculating the items' distance in psychological space. When *r* = 1 a city-block metric is used while when *r* = 2, a Euclidean distance metric is used. While $\boldsymbol{\omega}$ is often a free parameter fit to subject data, *r* is often defined by the type of stimuli used. When stimuli are highly discriminable and psychologically separable, a city-block metric is often used, while a Euclidian metric is used for integral dimensions that are not easily separable (Nosofsky, 2011).

**Influence of temporal context.** SAT-M includes an encoding strength (so named to differentiate them from the attention weight $\boldsymbol{\omega}$ defined above)**.** The critical parameter $\boldsymbol{\varepsilon}_i^x > 0$ in Equation 3 is the encoding strength of Feature *i* in Exemplar *x,* which depends on the sequence of study. SAT-M encodes items by comparing them to previously presented items. Differences and similarities between features from successive items and their category assignment define the strength of encoding of each feature $\boldsymbol{\varepsilon_i}$, i.e., the encoding weight given to that feature. Thus, for each dimension, the difference between *x* and the preceding item *y* is modulated by $\boldsymbol{\varepsilon}_i^x$ and its value depends on the match between the feature value in dimension *i* between *x* and *y* and their category assignments. The four possibilities are formally represented by four different encoding weights ($\varepsilon_{\text{different feature, same category}}$, $\varepsilon_{\text{different feature, different category}}$, $\varepsilon_{\text{same feature, same category}}$, $\varepsilon_{\text{same feature, different category}}$). Because the encoding weights depend on the properties of the current and previous stimuli and their category assignments, the same feature might (in fact, is likely to) receive different encoding weights from trial to trial and from participant to participant. Feature *i's* encoding weight is stored along with *x* during learning and modulates the influence that feature *i* has on the overall similarity metric and thus categorization.

The encoding weight assigned to a given feature is the result of the relation of that feature with those of the immediately preceding item and works to increase or decrease the importance of that feature for a stimulus as a function of sequence of study. Note that although we separate these four encoding weights, because there are four $\varepsilon_i$ free parameters fit to the data, it is possible that the value is the same for all four possibilities. This means that the differences in encoding between these four possibilities are not hard coded into the model, and the best fitting solution becomes an indicator for how important different features are as a function of their encoding context.

## Applications

We start by describing the target phenomenon we will use to demonstrate the role of local context in category learning and apply SAT-M. We will then show evidence that previous models cannot account for human behavior in such situations and demonstrate that SAT-M provides a good fit to the data, suggesting that local context – in addition to global context captured by other models – has an important influence on category learning. Finally, we will probe SAT-M's assumption that attention shifts on a local context, trial-by-trial, basis by using eye-tracking data to test whether the local attentional/encoding behavior of the model matches those of learners.

## Target phenomenon: The effect of sequence on category learning

In our application, we will focus on the behavioral results of one of the early studies on the impact that different sequences of study have on category acquisition: the results from Carvalho and Goldstone (2014). As described above, this phenomenon allows us to demonstrate the impact that local context changes have on category learning and how SAT-M captures this effect whereas models that do not include such a process of local attention changes do not closely match the results.

Carvalho and Goldstone (2014b) investigated whether different sequences of study would improve learning for different types of categories. Their main proposal was that different sequences of study lead to encoding different properties because learners' encoding of each of the items is dependent on what they have experienced before. Thus, if study involves repetition of the same category close in time, the encoding would emphasize what is similar among those items, whereas if study involves alternation between categories close in time, encoding would instead emphasize what is different among those items. This effect would be most salient when comparing learning of different types of categories. If the categories studied are highly *dissimilar*, noticing small *similarities* among items of the *same* categories – and by hypothesis blocked study -- should improve learning outcomes. Conversely, if the different categories studied are highly *similar*, noticing the small *differences* between items of *different* categories – and by hypothesis interleaved study – should improve learning outcomes.

To test these predictions, Carvalho and Goldstone (2014b) used two different sequences of study: blocked practice where items of a category are often followed by other items of the same category, and interleaved practice where items of one category are often followed by items of one of the other categories. They also manipulated the type of category being learned: either low similarity categories where items in the same category differed from each other and from items of other categories on most of their features, or high similarity categories where items of different categories differed from items of the same and different categories on only a small number of features.

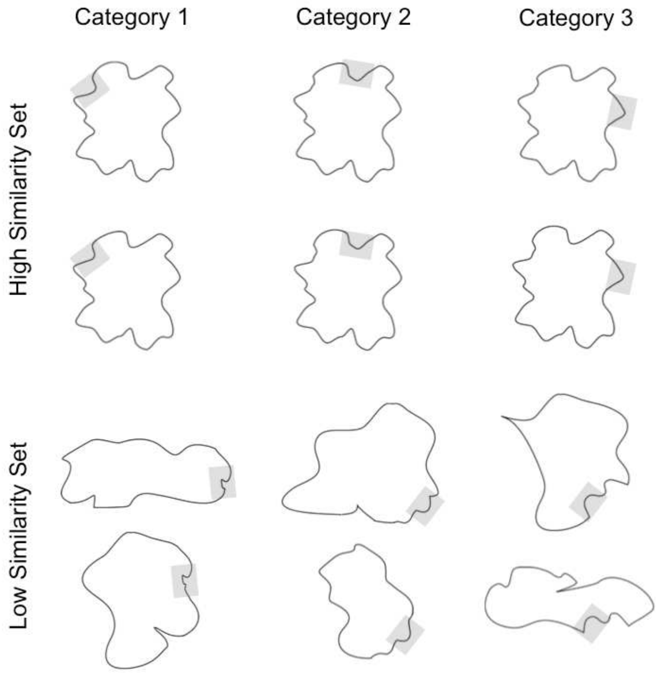

Figure 1: Example stimuli used in Carvalho & Goldstone (2014b). Stimuli in the high similarity set (top panel) differed on only two features between any two categories and only one feature among items of the same category. Stimuli in the low similarity set (bottom panel) differed in

many features between categories and among items of the same category. Grey boxes (not presented to participants) highlight the category-defining features). For details of the structure of the categories used see: https://osf.io/s87tf/.

The categories used were blob shapes composed of multiple curvilinear segments (see Figure 1 for examples). The structure of the categories is shown in Appendix A. Each dimension corresponds to a segment location in the blob whereas each feature value corresponds to a particular curvilinear segment. High and low similarity categories varied on how many differences existed between and within categories and thus on how many feature-values there were for each dimension (but not on the number of dimensions). All categories were defined by the presence of a particular curvilinear segment in a particular location of the space.

Participants started by studying three categories in one of the sequences and then being tested on those followed by another study phase in the other sequence and a new test phase. Each study phase was composed of 4 blocks of 72 trials each. During study 8 items of each category were presented 3 times each. The test phase included 48 trials, each a presentation of one of the 16 items of each category studied (half studied and half novel).

Overall, Carvalho and Goldstone (2014b) found that for high similarity categories, interleaved study improved classification of novel items at test, whereas for low similarity categories blocked study improved classification of novel items at test. This is the critical pattern of results that we will model in this paper.

Of course, one way to capture Carvalho and Goldstone's proposal and results described above would be to stipulate that the attentional weights for similarities and differences are different for blocked and interleaved sequences, for example, by fitting the two sequences separately. However, such an approach brings us back to the issue stated at the start of this paper. In so doing, one would be assuming that even though the sequence is a local change in context, it would be explained solely by global context, that is, by what category the item belongs to and the pressures that exerts on attention and encoding. Instead, we propose that SAT-M can parsimoniously account for the effect of different sequences on category learning and, moreover, better-characterize the flexibility of human learning.

## ALCOVE and SUSTAIN and categorization performance following different sequences

Our main proposal is that attention changes that occur over the course of category learning are the result not only of global context pressures arising from category classification, but also local context pressures arising from what other items were studied in closed temporal proximity. Importantly, previous models and theories of category learning fail to capture the effect of local context, focusing only on the global context.

To confirm this assertion, we fit two models that include an attention learning process during category learning: ALCOVE and SUSTAIN. Our point with these fits is to demonstrate that without considering the importance of local temporal context, these models cannot easily account for our previously presented category learning results, as we argued is the case when comparing the effect of different sequences of learning.

We fit both models to the target phenomenon described above. We used the implementations of ALCOVE and SUSTAIN available in Catlearn (Wills et al., 2017), an open archive of formal psychological models implemented in R (R Core Team, 2019). We trained the models with the study sequences from each participant and compared the model performance in the test phase with that of the participant trained with that sequence. In both models, category learning feedback was presented during training but not during test. We used sum of squared errors, implemented as part of the Catlearn package, as the objective function to determine best fitting parameters. The category representation described in Appendix A was adapted – maintaining its main characteristics -- to match how ALCOVE and SUSTAIN define category spaces.

ALCOVE has a total of 6 parameters. From these we defined the distance metric as Euclidian distance ($r$ = 2) and used a Gaussian similarity gradient ($q$ = 2). The remaining parameters were fit to the data. These parameters include the specificity constant ($c$), the decision constant ($\emptyset$) and the associative ($\lambda_w$) and attentional ($\lambda_\alpha$) learning rates. The best fitting values of these parameters are presented in Table 2.

SUSTAIN has a total of 5 parameters. From these, we defined the threshold to create a new cluster ($\tau$) as 0 because the study includes only supervised trials. We fit the following parameters: attentional focus ($r$), cluster competition ($\beta$), decision constancy ($d$), and learning rate ($\eta$). The best fitting parameter values are presented in Table 1. We defined the initial attentional weights and associative strengths as the same to all dimensions. Moreover, we defined the initial receptive fields as equally tuned and the network was always started with a single cluster with zero-strength weights.

Table 1: Best-fit parameter values for ALCOVE and SUSTAIN fits to Carvalho & Goldstone (2014b). Models were simulated using the implementation provided in the R package Catlearn.

| | ALCOVE | | | | SUSTAIN | | | |
|---|---|---|---|---|---|---|---|---|
| Parameter | $c$ | $\emptyset$ | $\lambda_w$ | $\lambda_\alpha$ | $r$ | $\beta$ | $d$ | $\eta$ |
| Best-fitting value | 15.85 | 1 | 0.718 | 0.99 | 11.99 | 0.99 | 1.13 | 0.49 |

The results of our simulations with best fitting parameters are shown in Figure 2. As can be seen in Figure 2, both models provide an overall poor fit to the data ($r^2 = .005$; *SSE* = 25.85, and $r^2 = .001$, *SSE*=31.54, for ALCOVE and SUSTAIN, respectively). Critically, both models also fail to show an interaction between category structure and sequence of study for novel items presented at test, a hallmark of the effect that local context has on category learning. The results of these simulations are consistent with our proposal that part of what is happening during category learning is learning to attend to information that is or is not relevant in the context of the other items studied in close temporal proximity (local context) and not only the global context of all the items belonging to either category. Although their specific mechanisms differ, attention in both ALCOVE and SUSTAIN changes during category learning as a result of the dimensions' general relevance for categorization. While both models are process models that take as input a specific sequence of trials, their dimension learning mechanisms are designed to work over long time courses, rather than the short trial-to-trial fluctuations of encoding that is implemented in SAT-M. One could argue that the poor fit of ALCOVE and SUSTAIN are in part the result of trying to fit both the high and low similarity categories simultaneously. One reason to fit each category structure separately would be to assume that each category structure constitutes a different global context where different information needs to be attended to and ignored. However, doing so would also assume that the cognitive system is "reset" to learn different structures in different ways and that it has a special way to tell which context it is while not specifying what that is. Similarly, one could argue that both ALCOVE and SUSTAIN could capture the results of Carvalho and Goldstone if fit separately to each sequence and category structure. Although that might be case, those fits do not answer the question of where learned attention to different features in different contexts originates. SAT-M, we propose, provides a parsimonious solution to this question.

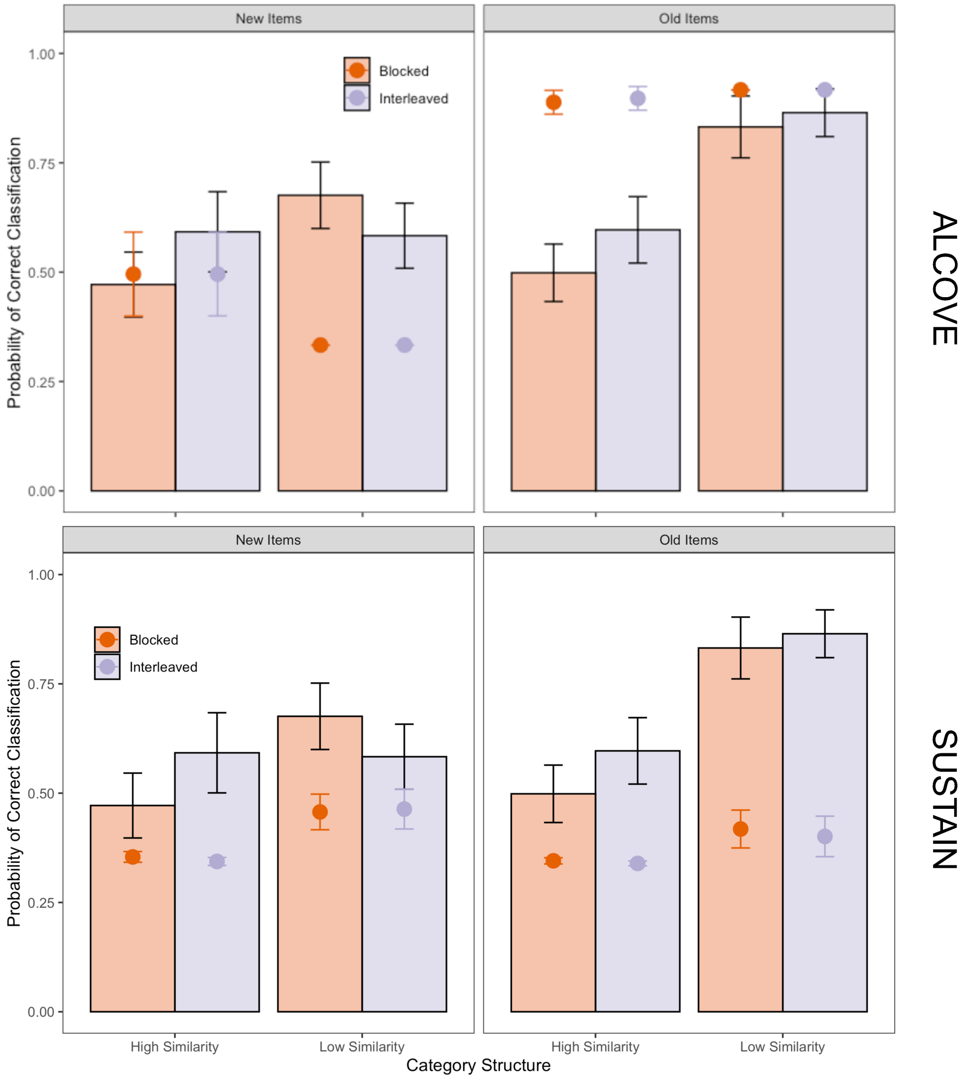

*Figure 2:* Fitting results (dots) for ALCOVE (top panel) and SUSTAIN (bottom panel) over the empirical results from Carvalho and Goldstone (2014b; represented by the bars). Best-fitting parameter values are presented in Table 1.

**SAT-M and categorization performance following different sequences**

In the previous section we demonstrated that extant models that include a learning process and a mechanism for attentional changes during learning cannot account for the empirical results reported by Carvalho and Goldstone (2014b). In this section, we explore whether the mechanism of sequential encoding of similarities and differences among successive items, when paired with different sequences of stimuli, can yield these results.  To that end, we fit SAT-M to the results from Carvalho and Goldstone (2014, Experiment 1). We fit SAT-M using maximum likelihood estimation with the four encoding weights ($\varepsilon_{\text{same feature, same category}}$, $\varepsilon_{\text{same feature, different category}}$, $\varepsilon_{\text{different feature, same category}}$, $\varepsilon_{\text{different feature, different category}}$), $\gamma$, and $c$ as free parameters fit to the data. We set $\rho =$ 2, because stimuli were relatively confusable and $r = 2$. We set $\omega = 1$ because the stimuli used were designed such that all dimensions were equally salient.

We fit the model using sequences generated in the same way as those shown to participants. For each run, the model was trained with a sequence similar to that of a human participant and then performance in a test phase including old and novel items was tested. In our simulations, each stimulus was represented as an 8-dimensional array. In the low similarity set, each dimension could assume a value 1-48 representing the many different components of that set of stimuli. For the high similarity set, each dimension could assume a value 1-4, representing the few differing components among stimuli. This representation matches how the stimuli were initially created (see Carvalho and Goldstone, 2014b, Appendix A, and materials stored in OSF: https://osf.io/s87tf/). The best-fitting $c$ and $\gamma$ parameters are shown in Table 2.

Table 2: Average best-fitting $c$ and $\gamma$ parameters for SAT-M fits and average best-fitting $c$, $\gamma$, *and* ε parameters SAT-M-R to results from Carvalho and Goldstone (2014, Experiment 1). Standard deviation in parenthesis.

| | *c* | *γ* | ε |
|---|---|---|---|
| SAT-M | 1.01 (0.46) | 4.12 (0.18) | -- |
| SAT-M-R | 1.00 (0) | 1.32 (0.00001) | 0.9 (0.00001) |

The results of the model are displayed as dots over the bars representing Carvalho and Goldstone's (2014b) results in the top panel of Figure 3. As it can be seen, the model provides a good fit to the data (*BIC* = 7053.60), suggesting that the effect of different sequences on category learning and generalization can be captured by a process of sequential comparison between successive items. For comparison, we constructed a restricted version of the model where the sequence of study does not influence the encoding weights of each stimulus property (Sequential Attention Theory Model Restricted; SAT-M-R). In SAT-M-R there is a single encoding weight $\boldsymbol{\varepsilon}$ that is a function of the feature's category relevance and does not vary depending on sequential variation among items (note that $\omega = 1$ as in SAT-M such that all dimensions are equally relevant for categorization). SAT-M-R embodies the plausible alternative proposal that the effect of different sequences – and local variation during study – can be well-accounted by global weighting based only on category assignment information. Best-fitting parameters are presented in Table 2.

As it can be seen in the bottom panel of Figure 3, SAT-M-R provides a poor fit to the data (*BIC* = 7319.03). Interestingly, although SAT-M-R displays the often-shown effect of stimulus similarity on categorization (better categorization of old items for low similarity categories, but better categorization of novel items for high similarity categories e.g., Palmeri & Nosofsky, 1995), it is not sensitive to the sequence of learning and cannot capture the interaction pattern evident in the human data. The Bayes factor comparing the two models directly suggests that SAT-M-R is much less likely to have generated the data than SAT-M ($B < 0.001$).

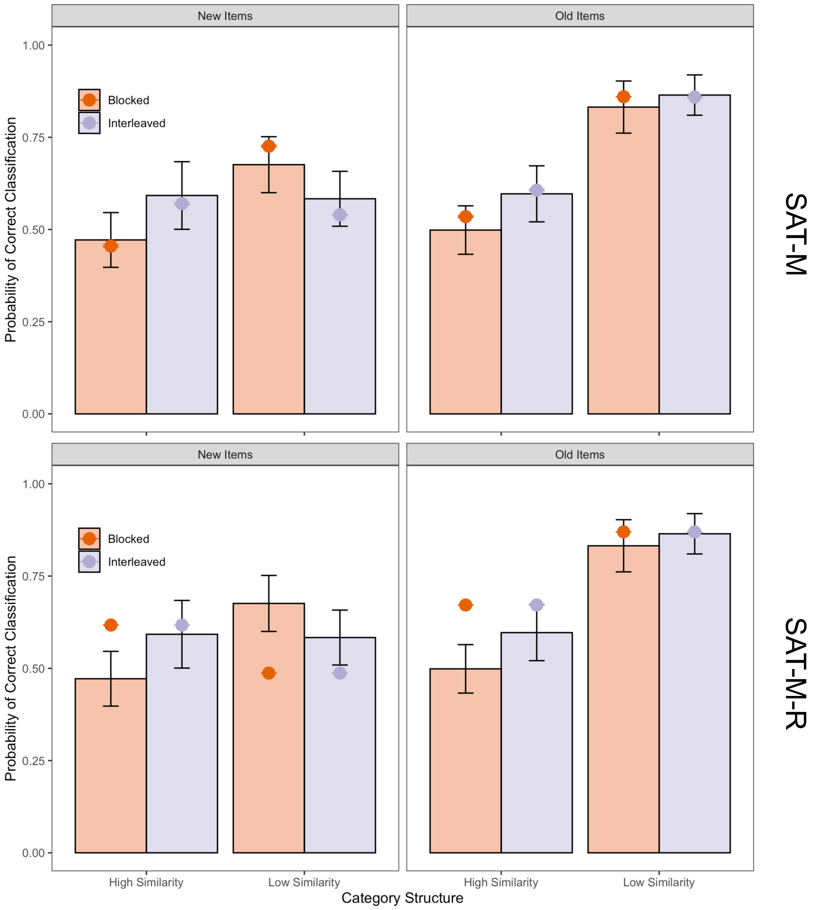

Figure 3. Fitting results (dots) for SAT-M (top panel) and SAT-M-R (bottom panel) over the empirical results from Carvalho and Goldstone (2014; represented by the bars). SAT-M provides a much better fit to the data than SAT-M-R.

The $\varepsilon$ parameters in SAT-M are sensitive to local context (i.e., the encoding parameters vary depending on the match in feature properties and category assignment across trials) and can be directly interpreted as the encoding strength of a feature. If encoding is sensitive to local context, then we should not only see better fit of SAT-M compared to a version in which ε is not allowed to vary due to local context as shown, but we should also see that the best-fit $\varepsilon$ values differ depending on local context. Importantly, in SAT-M, the encoding parameters are fit to the data and therefore the best fitting solution could be one where the encoding parameters, like is the case in SAT-M-R, are equivalent or vary in any number of ways. Table 1 shows the best fitting parameter values.

Table 3: Average best-fitting encoding parameters $\varepsilon$ when fitting SAT-M to results from Carvalho and Goldstone (2014, Experiment 1). Standard deviation in parenthesis.

| | Different Category | Same Category |
|---|---|---|
| Different Feature | 0.96 (0.08) | 0.0015 (0.01) |
| Same Feature | 0.11 (0.03) | 0.79 (0.22) |

As can be seen in Table 3, encoding strengths are stronger for differences between items of different categories and similarities among items of different categories. This result confirms that the SAT-M solution that best fits the data does so by varying encodings depending on local context (i.e., the match on category and feature properties between current and previous items). Moreover, these results are consistent with the theoretical explanation of the phenomenon offered by Carvalho and Goldstone (2014b): if interleaved study improves learning of high similarity categories because it emphasizes encoding of differences between categories whereas blocked study improves learning of low similarity categories because it emphasizes encoding of similarities among items of the same category, then we should see that the best fitting parameters should be higher for differences among different categories and similarities among items of the same category. Because different sequences include a disproportionate number of transitions across the same category (blocked study), or different categories (interleaved study), whether differences or similarities are better encoded depends on the sequence of study.

## Encoding weights and attention

The main novelty introduced by SAT-M is the sensitivity to local temporal context during category learning through differential encoding of each feature of a stimulus depending on how it compares to immediately preceding items and their category assignment. To this end, SAT-M introduces in the context of GCM the **ε** parameters that modulate encoding of a feature and vary depending on whether the feature and category assignment change relative to the previous item. Theoretically, we have argued that the inclusion of these parameters are sensible because learners are sensitive to local context in addition to global information about which dimensions are overall most relevant for category learning. Furthermore, we proposed that this sensitivity can be thought of as a modulation in how each stimulus experience is encoded and stored. The same item studied after a different previous item would be encoded differently.

In the previous section we showed that to fit data from an experiment showing how learners are sensitive to local context the **ε** parameters varied in their best-fitting values depending on whether the feature was the same or not and whether the category assignment was the same or not across two stimuli. If in fact local sensitivity can be captured by differential encoding patterns as in SAT-M, then we should see that looking patterns (often taken as a measure of overt attention), should be correlated to best-fit ε values.

Carvalho and Goldstone tested whether different sequences of study would lead learners to attend to different properties of the stimuli during different sequences of study. In their study participants studied the same two categories of Aliens either blocked or interleaved (see left panel of Figure 4 for examples of the stimuli used during study). The two categories studied had a structure such that multiple feature-values and dimensions could predict category assignment (discriminative features), and some features, despite being common in a category, did not predict category assignment (as they were also common in the other category – characteristic features; see Appendix B and stimuli available in https://osf.io/2n8gy/). After study, participants completed a test phase where participants classified novel items at test. The novel items could vary on the frequent but not discriminative (characteristic) feature or on other features (see right panel of Figure 4 for examples). Importantly, during learning participants eye movements were recorded using eye-tracking and how long learners spent looking at each of the items' features was analyzed.

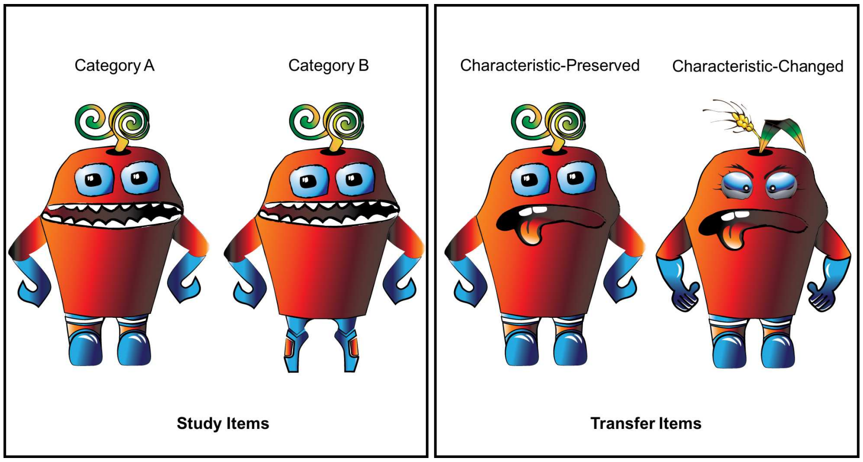

*Figure 4.* Example of stimuli from one of the families used by Carvalho and Goldstone (2017). Left Panel includes an example of each of the categories studied. Right panel includes an example of each of the novel items presented during the transfer task (both transfer items belong to Category A; equivalent items existed for category B). For details of the structure of the categories used see: https://osf.io/2n8gy/

Overall, the authors found that during interleaved study learners looked longer at properties that differed from those of the preceding item, whereas during blocked study learners attended to both differences and similarities equally (see left panel of Figure 5). The authors argued that the interaction is overall consistent with SAT in that following blocked study there is no bias towards looking at differences, even though those might be more salient because people often orient towards novelty or task-relevance (Rehder & Hoffman, 2005; Wang & Mitchell, 2011). Instead participants attend to repeated similarities to the same extent. During interleaved study, on the other hand, participants' attention is heavily biased towards differences between successive items, which are likely to be properties that differentiate between the categories.

Can these different attentional patterns be captured by SAT-M? If the encoding weights in SAT-M correspond to differential looking times as proposed, then the best-fitting parameters when we fit SAT-M to the categorization data should match the looking times observed in Carvalho and Goldstone (2017). To test these predictions, we fit SAT-M to the results of the testing phase of Experiment 3 of Carvalho and Goldstone (2017). Each stimulus used in the experiment was defined as a 5-dimensional array, with each dimension taking a value between 1 and 5. After fitting the model using the same parameters as described in the previous section, we extracted the best fitting encoding weight parameters for each feature of the studied items. The average best-fitting *c* parameter was 1.07 and the average best-fitting $\gamma$ parameter was 0.89. As described above, in SAT-M for each event the stimulus along with encoding parameters $\boldsymbol{\varepsilon}$ for each of its features are stored. As in the previous applications, SAT-M had four encoding parameters depending on the match of the properties and category assignment among the currently encoded item and the previous one – $\varepsilon_{\text{different feature, same category}}$, $\varepsilon_{\text{different feature, different category}}$, $\varepsilon_{\text{same feature, same category}}$, $\varepsilon_{\text{same feature, different category}}$.After fitting SAT-M to the behavioral results of the transfer task, we calculated, separately for the blocked and interleaved sequences, the sum across trials and features of the encoding parameters values for features that differed across successive items and features that were the same among successive items. Conceptually, summing up the encoding parameters is similar to summing up total looking time for each feature based on sequential differences and similarities, as done by Carvalho and Goldstone (2017).

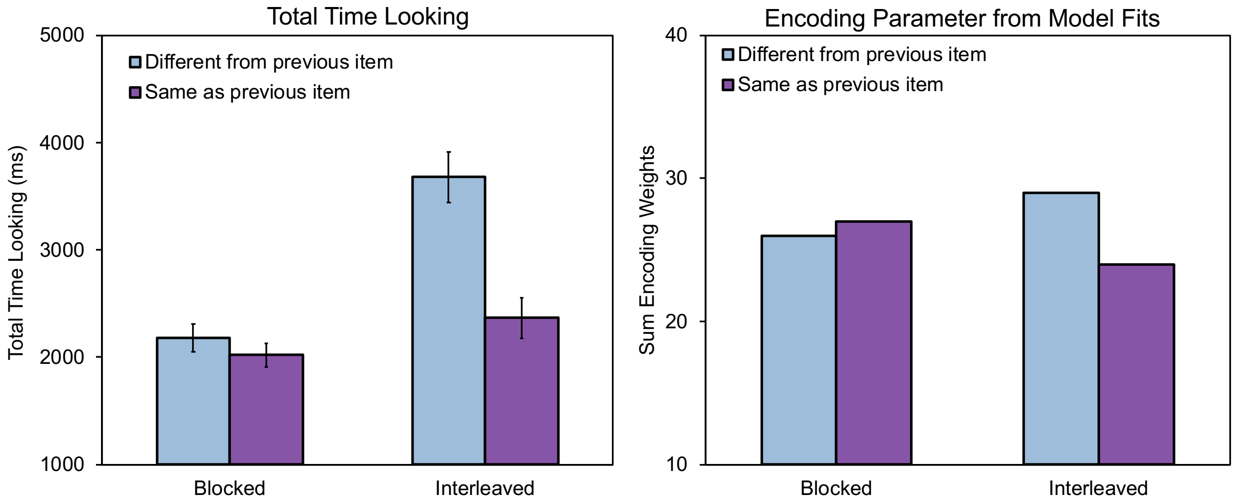

Figure 5: Results from total looking time during study in Carvalho and Goldstone's (2017) Experiment 3 (Left Panel) and summed best-fit values for encoding weights (ε) when SAT-M is fit to the categorization results of Carvalho and Goldstone's Experiment 3 (Right Panel).

As it can be seen in the right panel of Figure 5, for blocked study, the summed best-fitting encoding weight parameters are similar for differences and similarities across successive items. For interleaved study, on the other hand, the summed encoding weights are higher for differences than similarities across successive items. This pattern of results is similar to the empirical results found in the eye-tracking study (see left panel of Figure 5; Carvalho & Goldstone, 2017). It is also consistent with evidence from other eye-tracking studies investigating the influence of study sequence on category learning (Zaki & Salmi, 2019).

Overall, the results from these simulations suggest that the encoding weights in the SAT-M can be directly related to encoding differences in behavioral data. Importantly, we did not fit the encoding weights to the looking data in Carvalho and Goldstone (2017). Instead, we fit the model to the categorization test results and used the resulting best-fitting encoding parameter values as a measure of the model's encoding or looking time. Finding that when fit to the test data the model converges on encoding weights with a similar pattern to that of human looking time in the same paradigm is striking and speaks to the adequacy of the model's processes to those of human cognition.

## Implications

The flexibility of current models of categorization lies in large part on their ability to selectively weight some features more than others during learning. This procedure – although often successful at capturing how different features are differently relevant for different categories – fails to account for the full flexibility of human categorization. Beyond considering global variables such as the relevance of a property for correct classification, humans are sensitive to the local context of categorization and the same feature can become more or less relevant depending on the other items studied in close proximity. We developed a more flexible model of categorization – the Sequential Attention Theory Model (SAT-M), based on Carvalho and Goldstone's Sequential Attention Theory of category learning (Carvalho & Goldstone, 2017). The main assumption of our model is that how each item is encoded during category learning is not only the result of the global context of the categorization task, but also the local context of the temporally immediately preceding item. This assumption makes SAT-M not only flexible but also computationally efficient: in order to selectively encode properties that will tend to be useful for categorization SAT-M takes only the preceding item during learning. Thus, SAT-M requires very few items that need to be stored while relying mostly on the item that is expected to be the most accessible due to its recency.

To test SAT-M we compared its predictions with human behavior in a situation where local context strongly affects performance – the effect that difference sequences of items have on category learning. We have proposed that during category learning people engage in a process of sequential comparison to decide what is relevant and should therefore be encoded and what is not relevant and can be ignored. Engaging in this process, we proposed, is the reason why the sequence of study changes what is learned – different sequences create different sequential statistics that create attentional patterns towards encoding different types of features. The model introduced in this paper instantiated this proposal by including an initial sequential encoding process during which different features are assigned encoding weights depending on their relation to the features in the item encoded immediately before and whether both items belong to the same category or not. These encoding weights represent the likelihood of encoding a particular feature during a particular presentation of the item.

Importantly, these weights do not reflect the overall relevance of the feature for categorization as do the attention parameters in models such as GCM, ALCOVE, or SUSTAIN. A similarity shared by two successive items of the same category does not necessarily mean that it is diagnostic of categorization. Similarly, a difference between two successive items of different categories does not guarantee its diagnosticity for categorization. In this way, the sequence of study is biasing local encoding on a trial-by-trial level, as opposed to global attention to relevant vs. irrelevant properties at a task level. By strongly encoding locally relevant similarities that are not globally relevant (e.g., irrelevant within-category similarities in the high similarity category set), study in a blocked sequence might, at times, deter learning. Similarly, by encoding locally relevant differences that are not globally relevant (e.g., irrelevant between-category differences in the low similarity category set), interleaved study might, at times, deter learning. The opposite is equally important – encoding locally relevant similarities or differences that are relevant for categorization might boost performance because it does not depend on having experience with the full set of items and therefore knowing what has been relevant so far. Locally determined encodings could be seen as a catalyst for learning when a learner has not yet collected a lot of information yet as to the global relevance of different dimensions.

Notwithstanding the significance of attending to and encoding locally relevant features, the importance of global allocation of attention to category-relevant dimensions and away from category-irrelevant ones should not be ignored. There is widespread behavioral and eye tracking evidence that people learn to attend to certain dimensions while ignoring others, depending on their global predictive power for correct categorization (e.g., Blair et al., 2009; Chen et al., 2012; Rehder & Hoffman, 2005). In the current modeling we have focused only on local allocation of attention and encoding for simplicity. However, this means that because each item is encoded relative to only the previous item, dimensions that do not offer any predictive value continue to be attended. How could our model be extended to account for both local and global attention modulation? One possibility is to use the global attention parameter from GCM ($\omega$) to model how attention is modulated by the relevance of each dimension or feature in the context of all category items. However, this approach sidesteps understanding how global context affects attentional patterns during categorization. Another possibility would be to include global attention as a learned aspect of categorization and have a process of attention accumulation for each feature starting at the beginning of encoding, based on the geometric average of the encoding weights assigned for that dimension up to that point. In essence, high global attention to a dimension would be the result of continued high local attention and would reflect an inference of "If this dimension keeps being relevant locally, it must be important globally." With sufficient exposure, this global attention accumulator could then be used to modulate future local encoding decisions: A locally relevant difference between two categories on a seldom-predictive dimension might not be attended or encoded, while a locally irrelevant similarity between two categories on a frequently predictive dimension might be encoded. One potential advantage of this process compared to current models of categorization is that attention modulation would be a learned process at both the local (trial-by-trial) as well as the global (task) level.

In future extensions of this modeling work it would be important to include both local and global attention processes, and also to fit the data to individual participants' data by feeding the model with the sequence that a particular participant saw, using maximum-likelihood estimation of the best-fitting parameters for each participants' data. This would allow for a more finely tuned analyses of the differences between the encoding weights for different types of local changes and how this process affects global attention. Another possibility would be to allow earlier trials (beyond *N*-1 as used here) to affect the encoding weights. Although our decision to choose only the immediately preceding item is consistent with empirical evidence showing that people use the information from the previous trial heavily when making a new categorization decision (Jones & Sieck, 2003; Stewart et al., 2002), influences of preceding items could also be modeled by a function that decays more gradually (Stewart & Brown, 2004), perhaps varying with several factors such as how easy it is to encode different items (i.e., how confusable the exemplars are).

The results of the modeling work presented here have direct implications for theories of how the sequence of study influences learning by providing further evidence that a sequential comparison mechanism is plausible and can account for the results presented. We showed that there is no "best sequence" shortcut to learning. Instead, each sequence of study and the local attentional patterns it creates will shape what is learned. To understand whether a sequence is optimal or not one needs to understand what is encoded during learning and what is required during later training or transfer. A match will result in optimal learning. This proposal casts doubt over alternative proposals that focus on differentiation as the only basis for category learning and advocate for processes that maximize it (Kang & Pashler, 2012), or proposals arguing that the relative benefit of interleaved practice lies on the increased temporal spacing between items of the same category (Birnbaum et al., 2013; Kornell & Bjork, 2008) because it demonstrates that blocked study can result in best encoding of features repeated close in time and improved learning in situations that do not require discrimination.

More broadly, our work has implications for theories of category learning. Our model goes beyond conceptualizing attentional changes in category learning as the result of processing statistics over the entire course of study, that is, how the properties of a given item compare to the properties of all items of all categories. This is a powerful conceptual change from taking categorization as a broad learning process to a temporally local process, where each encoding moment is influenced by what just happened. Here we demonstrated that such a model can capture the effect of different sequences, but it might also capture other effects. Although typically described in the context of memory tasks, the spacing effect has also been shown in the context of category learning (e.g, Birnbaum et al., 2013). One way to conceptualize this effect in terms of changes in local context is to consider that encoding an item (or category of items) among many varied items increases encoding of different features because different features will be more relevant in pairwise comparisons, leading to a more robust encoding of the category and promoting later transfer.

More broadly, SAT-M and the importance of local context for attention modulation and encoding might also account for well-known effects in categorization. For example,  local attention changes can provide a mechanism through which category structure modulates category learning (Gureckis & Goldstone, 2008; Livingston & Andrews, 1995). By modulating attention on a trial-by-trial basis, it is likely that category boundaries would be emphasized even if they are not overall relevant for categorization. Similarly, the effect of local context can help explain the transfer of category learning across tasks (Gauthier et al., 2003), and the effect of different tasks (Trippas & Pachur, 2019) on category learning. These effects can be conceptualized in terms of changes in local context: when an item is learned among a certain type of items or semantic contexts, different properties will be emphasized that matter for that local context which can promote transfer (or hurt it).

## Appendix A

Table 1: Representation of the category structure used in Carvalho & Goldstone (2014b)

| D1 | D2 | D3 | D4 | D5 | D6 | D7 | D8 | Category | Similarity Structure | Item |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 2 | 2 | 2 | 2 | 2 | 1 | 1 | 1 | Low Similarity | LS_101 |
| 1 | 3 | 4 | 3 | 3 | 3 | 2 | 2 | 1 | Low Similarity | LS_102 |
| 1 | 4 | 5 | 4 | 4 | 4 | 3 | 3 | 1 | Low Similarity | LS_103 |
| 1 | 5 | 6 | 5 | 5 | 5 | 4 | 4 | 1 | Low Similarity | LS_104 |
| 1 | 6 | 7 | 6 | 6 | 6 | 5 | 5 | 1 | Low Similarity | LS_105 |
| 1 | 7 | 8 | 7 | 7 | 7 | 6 | 6 | 1 | Low Similarity | LS_106 |
| 1 | 8 | 9 | 8 | 8 | 8 | 7 | 7 | 1 | Low Similarity | LS_107 |
| 1 | 9 | 10 | 9 | 9 | 9 | 8 | 8 | 1 | Low Similarity | LS_108 |
| 1 | 10 | 11 | 10 | 10 | 10 | 9 | 9 | 1 | Low Similarity | LS_109 |
| 1 | 11 | 12 | 11 | 11 | 11 | 10 | 10 | 1 | Low Similarity | LS_110 |
| 1 | 12 | 13 | 12 | 12 | 12 | 11 | 11 | 1 | Low Similarity | LS_111 |
| 1 | 13 | 14 | 13 | 13 | 13 | 12 | 12 | 1 | Low Similarity | LS_112 |
| 1 | 14 | 15 | 14 | 14 | 14 | 13 | 13 | 1 | Low Similarity | LS_113 |
| 1 | 15 | 16 | 15 | 15 | 15 | 14 | 14 | 1 | Low Similarity | LS_114 |
| 1 | 16 | 17 | 16 | 16 | 16 | 15 | 15 | 1 | Low Similarity | LS_115 |
| 1 | 17 | 18 | 17 | 17 | 17 | 16 | 16 | 1 | Low Similarity | LS_116 |
| 2 | 1 | 19 | 18 | 18 | 18 | 17 | 17 | 2 | Low Similarity | LS_201 |
| 3 | 1 | 20 | 19 | 19 | 19 | 18 | 18 | 2 | Low Similarity | LS_202 |
| 4 | 1 | 21 | 20 | 20 | 20 | 19 | 19 | 2 | Low Similarity | LS_203 |
| 5 | 1 | 22 | 21 | 21 | 21 | 20 | 20 | 2 | Low Similarity | LS_204 |
| 6 | 1 | 23 | 22 | 22 | 22 | 21 | 21 | 2 | Low Similarity | LS_205 |
| 7 | 1 | 24 | 23 | 23 | 23 | 22 | 22 | 2 | Low Similarity | LS_206 |
| 8 | 1 | 25 | 24 | 24 | 24 | 23 | 23 | 2 | Low Similarity | LS_207 |
| 9 | 1 | 26 | 25 | 25 | 25 | 24 | 24 | 2 | Low Similarity | LS_208 |
| 10 | 1 | 27 | 26 | 26 | 26 | 25 | 25 | 2 | Low Similarity | LS_209 |
| 11 | 1 | 28 | 27 | 27 | 27 | 26 | 26 | 2 | Low Similarity | LS_210 |
| 12 | 1 | 29 | 28 | 28 | 28 | 27 | 27 | 2 | Low Similarity | LS_211 |
| 13 | 1 | 30 | 29 | 29 | 29 | 28 | 28 | 2 | Low Similarity | LS_212 |
| 14 | 1 | 31 | 30 | 30 | 30 | 29 | 29 | 2 | Low Similarity | LS_213 |
| 15 | 1 | 32 | 31 | 31 | 31 | 30 | 30 | 2 | Low Similarity | LS_214 |
| 16 | 1 | 33 | 32 | 32 | 32 | 31 | 31 | 2 | Low Similarity | LS_215 |
| 17 | 1 | 34 | 33 | 33 | 33 | 32 | 32 | 2 | Low Similarity | LS_216 |
| 18 | 18 | 1 | 34 | 34 | 34 | 33 | 33 | 3 | Low Similarity | LS_301 |
| 19 | 19 | 1 | 35 | 35 | 35 | 34 | 34 | 3 | Low Similarity | LS_302 |
| 20 | 20 | 1 | 36 | 36 | 36 | 35 | 35 | 3 | Low Similarity | LS_303 |

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| 21 | 21 | 1 | 37 | 37 | 37 | 36 | 36 | 3 | Low Similarity | LS_304 |
| 22 | 22 | 1 | 38 | 38 | 38 | 37 | 37 | 3 | Low Similarity | LS_305 |
| 23 | 23 | 1 | 39 | 39 | 39 | 38 | 38 | 3 | Low Similarity | LS_306 |
| 24 | 24 | 1 | 40 | 40 | 40 | 39 | 39 | 3 | Low Similarity | LS_307 |
| 25 | 25 | 1 | 41 | 41 | 41 | 40 | 40 | 3 | Low Similarity | LS_308 |
| 26 | 26 | 1 | 42 | 42 | 42 | 41 | 41 | 3 | Low Similarity | LS_309 |
| 27 | 27 | 1 | 43 | 43 | 43 | 42 | 42 | 3 | Low Similarity | LS_310 |
| 28 | 28 | 1 | 44 | 44 | 44 | 43 | 43 | 3 | Low Similarity | LS_311 |
| 29 | 29 | 1 | 45 | 45 | 45 | 44 | 44 | 3 | Low Similarity | LS_312 |
| 30 | 30 | 1 | 46 | 46 | 46 | 45 | 45 | 3 | Low Similarity | LS_313 |
| 31 | 31 | 1 | 47 | 47 | 47 | 46 | 46 | 3 | Low Similarity | LS_314 |
| 32 | 32 | 1 | 48 | 48 | 48 | 47 | 47 | 3 | Low Similarity | LS_315 |
| 33 | 33 | 1 | 49 | 49 | 49 | 48 | 48 | 3 | Low Similarity | LS_316 |
| 1 | 2 | 2 | 2 | 4 | 3 | 2 | 2 | 1 | High Similarity | HS_101 |
| 1 | 2 | 3 | 2 | 3 | 3 | 2 | 2 | 1 | High Similarity | HS_102 |
| 1 | 2 | 4 | 3 | 2 | 4 | 2 | 2 | 1 | High Similarity | HS_103 |
| 1 | 2 | 2 | 3 | 4 | 4 | 3 | 3 | 1 | High Similarity | HS_104 |
| 1 | 3 | 3 | 4 | 3 | 2 | 3 | 3 | 1 | High Similarity | HS_105 |
| 1 | 3 | 4 | 4 | 2 | 2 | 3 | 3 | 1 | High Similarity | HS_106 |
| 1 | 3 | 2 | 2 | 4 | 4 | 4 | 4 | 1 | High Similarity | HS_107 |
| 1 | 3 | 3 | 2 | 3 | 4 | 4 | 4 | 1 | High Similarity | HS_108 |
| 1 | 4 | 4 | 3 | 2 | 3 | 4 | 4 | 1 | High Similarity | HS_109 |
| 1 | 4 | 2 | 3 | 4 | 3 | 4 | 4 | 1 | High Similarity | HS_110 |
| 1 | 4 | 3 | 4 | 3 | 3 | 4 | 4 | 1 | High Similarity | HS_111 |
| 1 | 4 | 4 | 4 | 2 | 3 | 4 | 3 | 1 | High Similarity | HS_112 |
| 1 | 4 | 2 | 2 | 4 | 2 | 2 | 2 | 1 | High Similarity | HS_113 |
| 1 | 4 | 3 | 2 | 3 | 2 | 2 | 3 | 1 | High Similarity | HS_114 |
| 1 | 4 | 4 | 3 | 2 | 2 | 2 | 2 | 1 | High Similarity | HS_115 |
| 1 | 4 | 2 | 3 | 3 | 2 | 2 | 3 | 1 | High Similarity | HS_116 |
| 2 | 1 | 2 | 2 | 4 | 3 | 2 | 2 | 2 | High Similarity | HS_201 |
| 2 | 1 | 3 | 2 | 3 | 3 | 2 | 2 | 2 | High Similarity | HS_202 |
| 2 | 1 | 4 | 3 | 2 | 4 | 2 | 2 | 2 | High Similarity | HS_203 |
| 2 | 1 | 2 | 3 | 4 | 4 | 3 | 3 | 2 | High Similarity | HS_204 |
| 3 | 1 | 3 | 4 | 3 | 2 | 3 | 3 | 2 | High Similarity | HS_205 |
| 3 | 1 | 4 | 4 | 2 | 2 | 3 | 3 | 2 | High Similarity | HS_206 |
| 3 | 1 | 2 | 2 | 4 | 4 | 4 | 4 | 2 | High Similarity | HS_207 |
| 3 | 1 | 3 | 2 | 3 | 4 | 4 | 4 | 2 | High Similarity | HS_208 |
| 4 | 1 | 4 | 3 | 2 | 3 | 4 | 4 | 2 | High Similarity | HS_209 |
| 4 | 1 | 2 | 3 | 4 | 3 | 4 | 4 | 2 | High Similarity | HS_210 |
| 4 | 1 | 3 | 4 | 3 | 3 | 4 | 4 | 2 | High Similarity | HS_211 |

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| 4 | 1 | 4 | 4 | 2 | 3 | 4 | 3 | 2 | High Similarity | HS_212 |
| 4 | 1 | 2 | 2 | 4 | 2 | 2 | 2 | 2 | High Similarity | HS_213 |
| 4 | 1 | 3 | 2 | 3 | 2 | 2 | 3 | 2 | High Similarity | HS_214 |
| 4 | 1 | 4 | 3 | 2 | 2 | 2 | 2 | 2 | High Similarity | HS_215 |
| 4 | 1 | 2 | 3 | 3 | 2 | 2 | 3 | 2 | High Similarity | HS_216 |
| 2 | 2 | 1 | 2 | 4 | 3 | 2 | 2 | 3 | High Similarity | HS_301 |
| 2 | 3 | 1 | 2 | 3 | 3 | 2 | 2 | 3 | High Similarity | HS_302 |
| 2 | 4 | 1 | 3 | 2 | 4 | 2 | 2 | 3 | High Similarity | HS_303 |
| 2 | 2 | 1 | 3 | 4 | 4 | 3 | 3 | 3 | High Similarity | HS_304 |
| 3 | 3 | 1 | 4 | 3 | 2 | 3 | 3 | 3 | High Similarity | HS_305 |
| 3 | 4 | 1 | 4 | 2 | 2 | 3 | 3 | 3 | High Similarity | HS_306 |
| 3 | 2 | 1 | 2 | 4 | 4 | 4 | 4 | 3 | High Similarity | HS_307 |
| 3 | 3 | 1 | 2 | 3 | 4 | 4 | 4 | 3 | High Similarity | HS_308 |
| 4 | 4 | 1 | 3 | 2 | 3 | 4 | 4 | 3 | High Similarity | HS_309 |
| 4 | 2 | 1 | 3 | 4 | 3 | 4 | 4 | 3 | High Similarity | HS_310 |
| 4 | 3 | 1 | 4 | 3 | 3 | 4 | 4 | 3 | High Similarity | HS_311 |
| 4 | 4 | 1 | 4 | 2 | 3 | 4 | 3 | 3 | High Similarity | HS_312 |
| 4 | 2 | 1 | 2 | 4 | 2 | 2 | 2 | 3 | High Similarity | HS_313 |
| 4 | 3 | 1 | 2 | 3 | 2 | 2 | 3 | 3 | High Similarity | HS_314 |
| 4 | 4 | 1 | 3 | 2 | 2 | 2 | 2 | 3 | High Similarity | HS_315 |
| 4 | 2 | 1 | 3 | 3 | 2 | 2 | 3 | 3 | High Similarity | HS_316 |

*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). Each line represents a unique item.

## Appendix B

*Category structure for the stimuli used in Experiment 3 of Carvalho and Goldstone (2017).*

| Category | Item | Dimension 1 | Dimension 2 | Dimension 3 | Dimension 4 | Dimension 5 |
|---|---|---|---|---|---|---|
| A | 1 | 2 (1, 0.3) | 1 (0.5, 0.7) | 1 (0.5, 0.7) | 1 (0.5, 0.7) | 4 (0.5, 0.3) |
| A | 2 | 2 (1, 0.3) | 1 (0.5, 0.7) | 1 (0.5, 0.7) | 1 (0.5, 0.7) | 5 (0.5, 0.3) |
| A | 3 | 2 (1, 0.3) | 2 (1, 0.3) | 1 (0.5, 0.7) | 1 (0.5, 0.7) | 3 (0.5, 0.3) |
| A | 4 | 1 (0.5, 0.7) | 2 (1, 0.3) | 1 (0.5, 0.7) | 1 (0.5, 0.7) | 3 (0.5, 0.3) |
| A | 5 | 1 (0.5, 0.7) | 2 (1, 0.3) | 2 (1, 0.3) | 1 (0.5, 0.7) | 5 (0.5, 0.3) |
| A | 6 | 1 (0.5, 0.7) | 1 (0.5, 0.7) | 2 (1, 0.3) | 1 (0.5, 0.7) | 4 (0.5, 0.3) |
| A | 7 | 1 (0.5, 0.7) | 1 (0.5, 0.7) | 2 (1, 0.3) | 2 (1, 0.3) | 4 (0.5, 0.3) |
| A | 8 | 1 (0.5, 0.7) | 1 (0.5, 0.7) | 1 (0.5, 0.7) | 2 (1, 0.3) | 3 (0.5, 0.3) |
| A | 9 | 1 (0.5, 0.7) | 1 (0.5, 0.7) | 1 (0.5, 0.7) | 2 (1, 0.3) | 5 (0.5, 0.3) |
| B | 1 | 3 (1, 0.3) | 1 (0.5, 0.7) | 1 (0.5, 0.7) | 1 (0.5, 0.7) | 4 (0.5, 0.3) |
| B | 2 | 3 (1, 0.3) | 1 (0.5, 0.7) | 1 (0.5, 0.7) | 1 (0.5, 0.7) | 5 (0.5, 0.3) |
| B | 3 | 3 (1, 0.3) | 3 (1, 0.3) | 1 (0.5, 0.7) | 1 (0.5, 0.7) | 3 (0.5, 0.3) |
| B | 4 | 1 (0.5, 0.7) | 3 (1, 0.3) | 1 (0.5, 0.7) | 1 (0.5, 0.7) | 3 (0.5, 0.3) |
| B | 5 | 1 (0.5, 0.7) | 3 (1, 0.3) | 3 (1, 0.3) | 1 (0.5, 0.7) | 5 (0.5, 0.3) |
| B | 6 | 1 (0.5, 0.7) | 1 (0.5, 0.7) | 3 (1, 0.3) | 1 (0.5, 0.7) | 4 (0.5, 0.3) |
| B | 7 | 1 (0.5, 0.7) | 1 (0.5, 0.7) | 3 (1, 0.3) | 3 (1, 0.3) | 4 (0.5, 0.3) |
| B | 8 | 1 (0.5, 0.7) | 1 (0.5, 0.7) | 1 (0.5, 0.7) | 3 (1, 0.3) | 3 (0.5, 0.3) |
| B | 9 | 1 (0.5, 0.7) | 1 (0.5, 0.7) | 1 (0.5, 0.7) | 3 (1, 0.3) | 5 (0.5, 0.3) |

*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).

# References

Aha, D. W., & Goldstone, R. L. (1992). Concept learning and flexible weighting. *Proceedings of the Fourteenth Annual Conference of the Cognitive Science Society*, 534–539.

Anderson, J. R., Matessa, M., & Lebiere, C. (1997). ACT-R: A Theory of Higher Level Cognition and Its Relation to Visual Attention. *Human–Computer Interaction*, *12*(4), 439–462. https://doi.org/10/bmhrqt

Ashby, F. G., Paul, E. J., & Maddox, W. T. (2011). COVIS. In *Formal approaches in categorization* (pp. 65–87).

Bareiss, E. R., Porter, B. W., & Wier, C. C. (1990). PROTOS: An exemplar-based learning apprentice. In Y. Kodratoff & R. S. Michalski (Eds.), *Machine Learning* (pp. 112–127). Morgan Kaufmann. https://doi.org/10.1016/B978-0-08-051055-2.50009-2

Birnbaum, M. S., Kornell, N., Bjork, E. L., & Bjork, R. A. (2013). Why interleaving enhances inductive learning: The roles of discrimination and retrieval. *Memory & Cognition*, *41*(3), 392–402.

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*(5), 1196–1206.

Brunmair, M., & Richter, T. (in press). Similarity matters: A meta-analysis of interleaved learning and its moderators. *Psychological Bulletin*. https://doi.org/10.1037/bul0000209

Carvalho, P. F., & Goldstone, R. L. (2014a). Effects of interleaved and blocked study on delayed test of category learning generalization. *Frontiers in Psychology*, *5*, 1–11. https://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*(3), 481–495. https://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*(1), 281–288. https://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*(APR). https://doi.org/10.3389/fpsyg.2015.00505

Carvalho, P. F., & Goldstone, R. L. (2017). The sequence of study changes what information is attended to, encoded, and remembered during category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *43*(11), 1699–1719. https://doi.org/10.1037/xlm0000406

Chen, L., Meier, K. M., Blair, M. R., Watson, M. R., & Wood, M. J. (2012). Temporal characteristics of overt attentional behavior during category learning. *Attention, Perception, & Psychophysics*, *75*(2), 244–256.

Gauthier, I., James, W., Curby, K. M., & Tarr, M. J. (2003). The influence of conceptual knowledge on visual discrimination. *Cognitive Neuropsychology*, *20*(3), 507–523.

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

Gureckis, T. M., & Goldstone, R. L. (2008). The effect of the internal structure of categories on perception. In B. C. Love, K. McRae, & V. M. Sloutsky (Eds.), *Proceedings of the 30th*

*Annual Conference of the Cognitive Science Society* (pp. 1876–1881). Cognitive Science Society.

Hintzman, D. L. (1984). MINERVA 2: A simulation model of human memory. *Behavior Research Methods, Instruments, & Computers*, *16*(2), 96–101.

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*(2), 316–332. https://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*(4), 626–640. https://doi.org/10.1037/0278-7393.29.6.1118

Kang, S. H. K., & Pashler, H. (2012). Learning Painting Styles: Spacing is Advantageous when it Promotes Discriminative Contrast. *Applied Cognitive Psychology*, *26*, 97–103.

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

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

Livingston, K. R., & Andrews, J. K. (1995). On the Interaction of Prior Knowledge and Stimulus Structure in Category Learning. *The Quarterly Journal of Experimental Psychology Section A*, *48*(1), 208–236.

Love, B. C., Medin, D. L., & Gureckis, T. M. (2004). SUSTAIN: A Network Model of Category Learning. *Psychological Review*, *111*(2), 309–332.

Luce, R. D. (1963). Detection and recognition. In R. D. Luce, R. R. Bush, & E. Galanter (Eds.), *Handbook of Mathematical Psychology* (pp. 103–189). Wiley.

Nosofsky, R. M. (1986). Attention, similarity, and the identification-categorization relationship. *Journal of Experimental Psychology. General*, *115*(1), 39–61.

Nosofsky, R. M. (2011). The generalized context model: An exemplar model of classification. In E. M. Pothos & A. J. Wills (Eds.), *Fromal Approaches in Categorization* (pp. 18–39). Cambridge University Press.

Nosofsky, R. M., & Zaki, S. R. (2002). Exemplar and prototype models revisited: Response strategies, selective attention, and stimulus generalization. *Journal Of Experimental Psychology-Learning Memory And Cognition*, *28*(5), 924–940. https://doi.org/10.1037/0278-7393.28.5.924

Palmeri, T. J., & Nosofsky, R. (1995). Recognition memory for exceptions to the category rule. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *21*(3), 548–568.

R Core Team. (2019). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Rehder, B., & Hoffman, A. B. (2005). Thirty-something categorization results explained: Selective attention, eyetracking, and models of category learning. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *31*(5), 811–829. https://doi.org/10.1037/0278-7393.31.5.811

Schyns, P. G., & Rodet, L. (1997). Categorization creates functional features. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *23*(3), 681–696.

Shepard, R. N. (1965). Approximation to Uniform Gradients of Generalization by Monotone Transformation of Scale BT - Stimulus Generalization. In D. I. Mostofsky (Ed.), *Stimulus Generalization* (pp. 94–110). Stanford University Press.

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*(2), 416–430.

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*(1), 3–11.

Thiessen, E. D., & Pavlik, P. I. (2013). iMinerva: A Mathematical Model of Distributional Statistical Learning. *Cognitive Science*, *37*(2), 310–343. https://doi.org/10.1111/cogs.12011

Trippas, D., & Pachur, T. (2019). Nothing compares: Unraveling learning task effects in judgment and categorization. *Journal of Experimental Psychology: Learning, Memory, and Cognition*. https://doi.org/10/gf5sr4

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

Wang, T., & Mitchell, C. J. (2011). Attention and relative novelty in human perceptual learning. *Journal of Experimental Psychology: Animal Behavior Processes*.

Wills, A. J., O'Connell, G., Edmunds, C. E. R., & Inkster, A. B. (2017). Progress in Modeling Through Distributed Collaboration. In *Psychology of Learning and Motivation* (Vol. 66, pp. 79–115). Elsevier. https://doi.org/10.1016/bs.plm.2016.11.007

Zaki, S. R., & Salmi, I. L. (2019). Sequence as context in category learning: An eyetracking study. *Journal of Experimental Psychology: Learning, Memory, and Cognition*. https://doi.org/10/gf6rw9