The Dynamics of Exemplar and Prototype Representations Depend on Environmental Statistics

Arjun Devraj (adevraj@princeton.edu)
Department of Computer Science, Princeton University

Qiong Zhang (qiongz@princeton.edu)
Princeton Neuroscience Institute, Princeton University

Thomas L. Griffiths (tomg@princeton.edu)
Departments of Psychology and Computer Science, Princeton University

Abstract

How people represent categories—and how those representations change over time—is a basic question about human cognition. Previous research has suggested that people categorize objects by comparing them to category prototypes in early stages of learning but use strategies that consider the individual exemplars within each category in later stages. However, many category learning experiments do not accurately reflect the environmental statistics of the real world, where the probability that we encounter an object changes over time. Our goal in this study was to introduce memory constraints by presenting each stimulus at intervals corresponding to the power-law function of memory decay. Since the exemplar model relies on the individual's ability to store and retrieve previously seen exemplars, we hypothesized that adding memory constraints that better reflect real environments would favor the exemplar model more early on compared with later. Confirming our hypothesis, the results illustrate that under realistic environmental statistics with memory constraints, the exemplar model's advantage over the the prototype model decreases over time.

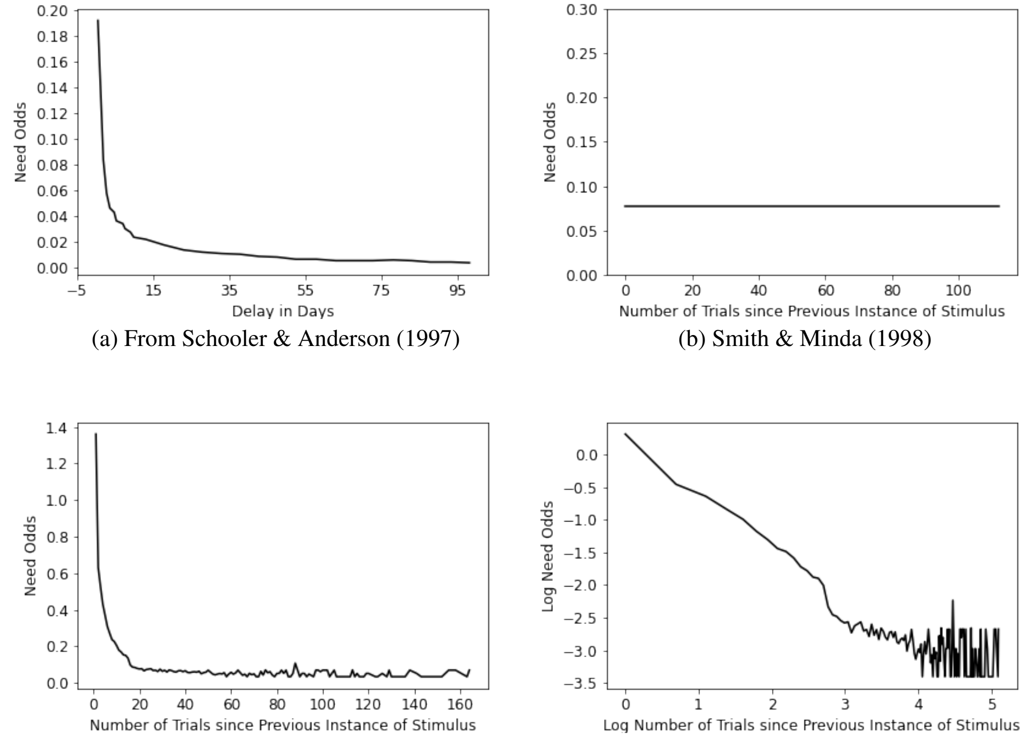
Keywords: category learning; prototype model; exemplar model; memory decay; environmental statistics

Introduction

Computational models of categorization have played a key role in understanding the representations people use when learning categories. Two prominent models of categorization include the prototype model, which posits that a stimulus is categorized by comparing it to solely the "mean" object, or prototype, of each category (Posner & Keele, 1968; Reed, 1972), and the exemplar model, which asserts that a stimulus is categorized by comparing it with all of the objects for each category (Medin & Schaffer, 1978). Earlier findings strongly favored the exemplar model (McKninley & Nosofsky, 1996; Medin & Schaffer, 1978; Shin & Nosofsky, 1992), but these did not consider the progression of category learning over time. Smith & Minda (1998) studied the fits of the prototype and exemplar models of categorization to human performance on a word categorization task over time and found that although the exemplar model dominates in the end, the prototype model has a strong advantage in early stages of learning, particularly with non-linearly separable (NLS) category structures. This result is consistent with the proposal that individuals categorize using a simple rule-based approach in early epochs and more exemplar-specific strategies in later epochs (Nosofsky et al., 1994).

While prototype-based categorization vs. exemplar-based categorization has been extensively studied in the literature, less attention has been paid to the constraints of memory on categorization; namely, how much one is able to utilize prototype-based or exemplar-based categorization is constrained by how accessible these representations are in their memory. Prototype-based categorization requires retrieving generalized knowledge about the category, whereas exemplar-based categorization requires retrieving memory about specific instances. It has been shown that one categorizes using generalized knowledge in early epochs and switches to an exemplar-based model in later stages (Smith & Minda, 1998), which does not seem consistent with findings in the memory literature showing that it becomes increasingly easier to access representations of general knowledge than representations of specific items over time—i.e., memory of specific items decays much faster than memory of the gist, or general knowledge (Posner & Keele, 1970; Zeng et al., 2021). How would one rely more on exemplar-based representations in late stages of categorization when it is difficult to access these exemplars in memory?

Memory of specific items decays quickly for a reason. From a rational perspective, there is no need to retain information in memory if it is no longer needed, as determined by the statistics of the environment (Anderson & Schooler, 1991). Therefore, memory decay obeys a power law function because the relationship between stimulus recency and the need odds, the odds that the stimulus will be required in the future, often obeys the power law in realistic environments, as examined from environmental sources such as *The New York Times*, parental speech, and electronic mail (Schooler & Anderson, 1997; Figure 1a). In contrast, categorization experiments, including Smith & Minda (1998), usually use stimuli that appear with the same frequency regardless of the last time the stimulus was seen (Figure 1b). This provides an opportunity for overlearning individual stimuli, which is unrealistic given the environmental statistics encountered in the real world (Figure 1a). To fill this gap, we designed an experiment that better reflects the power-law relationship between stimulus recency and the need odds (Figures 1c, 1d). We hypothesize that setting up the categorization experiment with realistic environmental statistics will reduce the accessibility of exemplar-based representations in the memory over time, therefore reducing or inverting the trend previously observed in favoring prototype-based representation early on



(a) From Schooler & Anderson (1997)

(b) Smith & Minda (1998)

(c) Our Experiment

(d) Our Experiment, in Log-Log Coordinates

Figure 1: The environmental recency function, displaying the relationship between stimulus recency and its need odds, calculated as $P(\text{stim})/(1-P(\text{stim}))$ where $P(\text{stim})$ is the probability of seeing the stimulus. (a) is based on New York Times headlines (1986 and 1987), reproduced from Figure 4a in Schooler & Anderson (1997). (b) is calculated based on the experimental design of Smith & Minda (1998). (c) and (d) display average data from the 60 sequences of stimuli generated for the experimental condition of our experiment. In (c) and (d), $P(\text{stim})$ was calculated using a window size of 15 trials. The linear log-log relationship in (d) indicates that (c) is a power-law function.

and exemplar-based representations at late stages.

We expand on the work of Smith & Minda (1998) using Anderson & Schooler's (1991) observations about realistic environmental structures and memory retention to redesign the experiment. In the control condition, we replicated Smith & Minda's second experiment (with the NLS category structures), in which stimuli are presented with uniform frequency over time. In the experimental condition, however, we constrained the presentation of stimuli to reflect the power-law function in order to better simulate a real-world environment involving memory constraints (Anderson & Schooler, 1991). Since the exemplar model requires the storage and retrieval of all exemplars, later trials would involve heightened memory constraints.

The plan of the paper is as follows. We first explain the formulations of the prototype and exemplar models of categorization. Next, we delineate how we constructed the presentation of stimuli in the experimental condition in order to reflect the power-law model encountered in the real world. Finally, we discuss the behavioral and model results from both our replication of Smith & Minda (1998) in the control condition and our manipulations of the environmental statistics in the experimental condition.

Background

Prototype Model

Various versions of the prototype model exist (Medin & Schaffer, 1978; Medin & Smith, 1981; Reed, 1972); we use here the formulation specified by Smith & Minda (1998, 2011). The prototype model compares the observed stimulus to the prototype for each category. First, the distance d_{i,P_k} between stimulus i and the prototype for category $k \in \{1,2\}$, P_k, is computed as

$$d_{i,P_k} = \sum_{j=1}^{m} w_j |i_j - P_{k,j}|, \tag{1}$$

where i_j is the stimulus' value for dimension j, and $P_{k,j}$ is the the value of the prototype of category k for dimension j, and w_j is the attentional weight assigned to dimension j.

Each attentional weight w_j is constrained to take on a value in $[0,1]$, and the weights together sum to 1 ($\sum_{j=1}^{m} w_j = 1$).

Once the raw distance d_{i,P_k} has been calculated, stimulus i's similarity, η_{i,P_k}, to the category k prototype P_k is computed as

$$\eta_{i,P_k} = e^{-c*d_{i,P_k}}, \tag{2}$$

where c is a sensitivity parameter constrained to $[0,20]$. The sensitivity parameter has the effect of amplifying or shrinking psychological space (Smith & Minda, 1998). Finally, the probability that stimulus i (S_i) will be categorized into category 1 (R_1) is given by

$$P(R_1|S_i) = \frac{\eta_{i,P_1}}{\eta_{i,P_1} + \eta_{i,P_2}}. \tag{3}$$

Exemplar Model

The exemplar model has been developed and generalized from the original context model (Medin, 1975; Nofosky 1984, 1986, 1987, 1988; Palmeri & Nofosky, 1995; McKinley & Nosofsky, 1995). This formulation of the exemplar model was used and thoroughly described in Smith & Minda (1998). The exemplar model compares the observed stimulus to all of the previously seen exemplars in each category in order to generate a category prediction for the observed stimulus. The distance $d_{i,x}$ and similarity measure $\eta_{i,x}$ between stimulus i and exemplar x are calculated identically as in the prototype model. However, the exemplar model considers all exemplars in order to generate a prediction, so the probability that stimulus i (S_i) will be categorized into category 1 (R_1) is

$$P(R_1|S_i) = \frac{\sum_{x \in C_1} \eta_{i,x}}{\sum_{x \in C_1} \eta_{i,x} + \sum_{x \in C_2} \eta_{i,x}}, \tag{4}$$

where C_1 is the set of exemplars for category 1 and C_2 is the set of exemplars for category 2.

While Nosofsky & Zaki (2002) criticized the lack of a response-scaling parameter in the exemplar model used in Smith & Minda (1998), extensive debate in the literature (Myung, Pitt, & Navarro, 2007; Smith & Minda, 1998, 2002) led us to exclude this parameter in order to reduce the complexity of the model and enable direct comparison to the original experiment. Because our experiment used stimuli with six dimensions, both models had a total of seven parameters: the six attentional weights ($w_1, ..., w_6$) and the sensitivity parameter (c). The two models were then fit to participant data from the control and experimental conditions of the experiment.

Methods

Our experiment aims to replicate Experiment 2 (NLS categories) from Smith & Minda (1998) in the control condition and made modifications to introduce memory constraints in the experimental condition. The task required participants to repeatedly categorize 14 stimuli into two categories. The stimuli were six-letter nonsensical words taken from Appendix A of Smith & Minda (1998). Each stimulus can be thought of as a six digit string of bits, such that the prototype for category 1 was `000000` and the prototype for category 2 was `111111`. Seven stimuli belonged to category 1 [`000000`, `100000`, `010000`, `001000`, `000010`, `000001`, `111101`], and the other seven stimuli belonged to category 2 [`111111`, `011111`, `101111`, `110111`, `111011`, `111110`, `000100`]. Each digit and position in the binary string corresponds to a unique letter. In each trial, the participant was shown one of the 14 stimuli and had unlimited time to select a response of *1* for category 1 and *2* for category 2. For exactly one second after each trial, the participant was shown *Correct* or *Incorrect* accordingly.

We collected data from 60 participants for each condition. (For reference, Smith & Minda (1998) had 32 participants.) Participants were recruited from the 18-23 age range and English-speaking population using Prolific. The hypothesis, design, sample size, and analysis of this experiment have been preregistered (`https://aspredicted.org/yq984.pdf`). Scripts for the exemplar and prototype models were derived from the open source code (`https://github.com/jpminda/Categorization_Models`).

Control Condition

The control condition, similar to Smith & Minda's (1998) experiment (the only difference being 616 trials in our experiment vs. 560 trials in the original experiment), consisted of 616 trials divided into 44 blocks, each consisting of 14 trials. Each block consisted of a random permutation of the 14 stimuli such that each stimulus was shown exactly once per block.

Experimental Condition

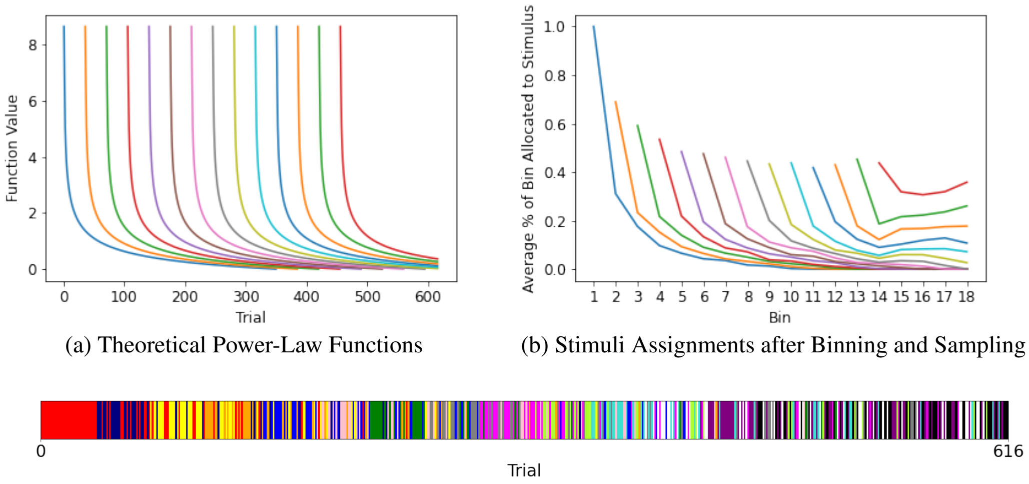
To create the environmental statistics in Figure 1a, we developed an algorithm to generate a sequence of stimuli for the experiment. The specific pattern of stimulus presentation should, in turn, influence the need odds and hence the participant's retention of objects in memory. The experimental design first involved assigning each of the fourteen stimuli its own power-law function, generically defined as

$$f(x) = ax^k. \tag{5}$$

We added several new parameters to the power-law function to generate the necessary experimental conditions:

$$f(t) = \begin{cases} 0 & t < t_0 \\ s & t = t_0 \\ a(\frac{t-t_0}{r})^k - \theta & t_0 < t \le t_0 + n \\ 0 & t > t_0 + n \end{cases} \tag{6}$$

where t is the trial number, t_0 is the start trial for the stimulus to be introduced, r is the range, n is the number of trials the object's power-law function should be sampled from, s is the starting value for the power-law function, and θ is a calibration parameter. Power-law functions for consecutive stimuli were introduced in intervals of 35 trials. As in the control condition, each participant participated in a total of 616 trials. Figure 2a displays the resulting power-law functions for



(a) Theoretical Power-Law Functions

(b) Stimuli Assignments after Binning and Sampling

(c) Example Stimulus Sequence for a Single Participant

Figure 2: (a) Each curve corresponds to the power-law function for a particular stimulus. (b) Since data is aggregated, each curve corresponds to the i-th stimulus seen across all participants. For example, the first curve illustrates the average proportion of each bin allocated to the first stimulus (regardless of its identity) shown to the 60 participants. (c) Each color corresponds to a unique stimulus $i \in \{1, 14\}$, and the figure highlights the significant temporal clustering and subsequent decay caused by our stimulus generation algorithm based on power-law functions. Parameter values of $a = 1$, $k = -0.3$, $r = 1000$, $n = 350$, $s = 10$, and $\theta = a * (\frac{n}{r})^k$ were chosen to adequately reflected the power-law function in Figure 1a.

each stimulus over the entirety of the experiment. The order of the stimuli was randomly assigned for each participant.

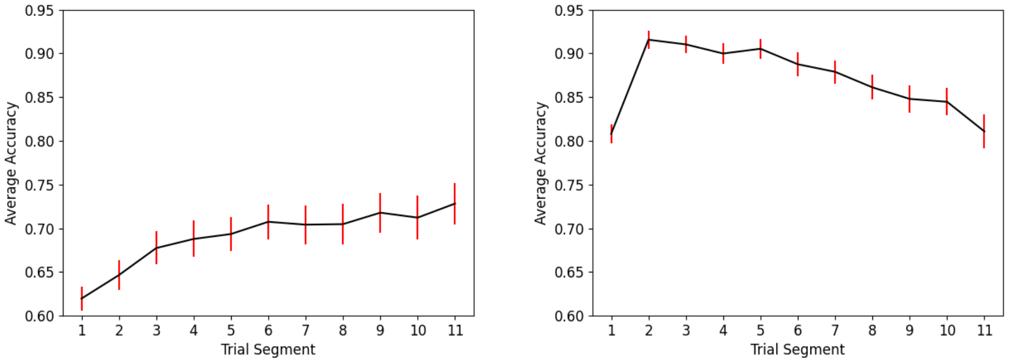
To assign a stimulus to each trial based on the continuous power-law functions, we binned trials into discrete bins of 35 trials, generated a probability distribution for the bin based on the relative values of the various stimuli power-law functions over the trials in the bin, and generated the stimuli for trials in the bin by sampling from this distribution. Figure 2b shows how the resulting sequences of stimuli adequately reflected the original power-law functions. Figure 2c shows an example of stimuli sequence for a single participant. Finally, we verified that the sequences of stimuli generated by this method reflected the environmental statistics in Figure 1a.

Model Fitting

Trials were split and grouped into 11 trial segments, each containing $\frac{1}{11}$th of the trials for each stimulus based on when the trial occurred in the experiment. Conceptually, for the experimental condition, the first trial segment corresponds to the earliest point along all stimuli power-law functions (highest retention), and the last trial segment corresponds to the latest point (lowest retention); for the control condition, trial segments correspond to chronological groupings of trials in absolute time, exactly as in Smith & Minda (1998).

For the comparison of the control condition to Smith and Minda's (1998) results, we used the sum of squared errors (SSE) between the observed and predicted probabilities as the measure of fit, which we sought to minimize. We ran Scipy's Sequential Least Squares Programming (SLSQP) method with 10 initial random configurations to obtain the best-fit parameters. The model was fit for each participant receiving the control condition over each trial segment, as in Smith & Minda (1998). The resulting best-fit parameters were used to calculate the SSE that was ultimately used as the measure of fit in Figure 4.

However, the same model-fitting procedure and definition of fit are not germane for the experimental condition because trial segments did not represent chronological groupings of trials in absolute time; consequently, we could not use the models to generate predicted probabilities for entire trial segments, as these contained many trials *distributed* over time, and not the exact sequences that were seen by the participant. (For the exemplar model, this is especially problematic because each trial within a given trial segment was seen at a different point in time, thereby involving a different number of previously seen exemplars. Any probabilities generated by the model for an entire trial segment could not possibly account for these differences.) Hence, we used the mean squared error (MSE) within a trial segment, calculated on a trial-by-trial basis (automatically setting the predicted probabilities of trial-irrelevant stimuli to zero), as the measure of fit, and fit the model for each participant over the entire experiment (rather than per trial segment) using the SLSQP method with 10 initial random configurations. We ensured that the exemplar model only considered exemplars seen *thus far* since not all exemplars were seen from the beginning. The new model-fitting procedure and MSE fit calculation were neces-



(a) Control Condition

(b) Experimental Condition

Figure 3: Average categorization accuracy over trial segments. Error bars denote the standard error of the mean.

sary so that we could apply the models to an individual trials (since it would be inappropriate to apply the model to non-contiguous sets of trials constituting a trial segment and generate meaningless predicted probabilities based on a set of data that was not actually observed by the participant, particularly when different trials within a given trial segment might involve a different set of previously seen exemplars) and then average the squared per-trial residuals to derive a metric of the "fit" for a given trial segment. For ease of comparison, the same model-fitting procedure was applied to the control condition, and the comparison is shown in Figure 5.

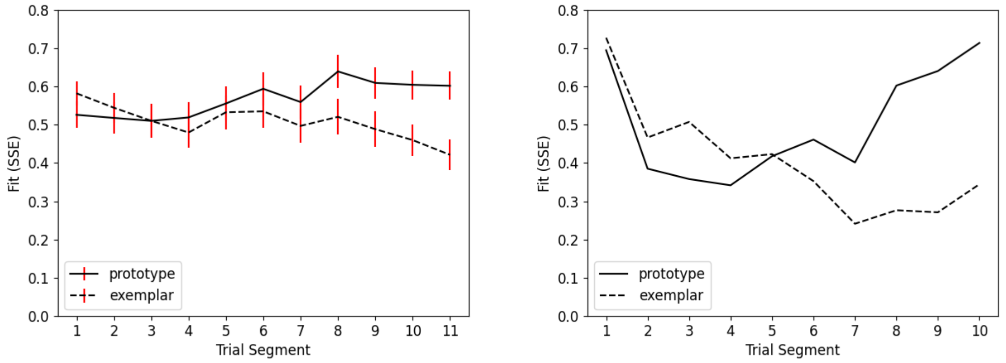
Results

Behavioral Results

The categorization accuracies per trial segment in each condition were averaged over all participants and graphed over time, resulting in Figure 3. In the control condition, significant learning occurred between the first (M =.62, SD =.11) and last (M =.73, SD =.18) trial segment, $t(59) = -5.08$, $p < .001$. In the experimental condition, memory constraints effectively caused participant classification accuracy to decrease from the second (M =.92, SD =.08) to last (M =.81, SD =.15) trial segment, $t(59) = 7.99$, $p < .001$. The second trial segment was compared in the experimental condition because this was the point that best reflected the beginning of the power-law function—when the participant has repeatedly seen the stimulus and is past the very initial stage of learning.

Model Results

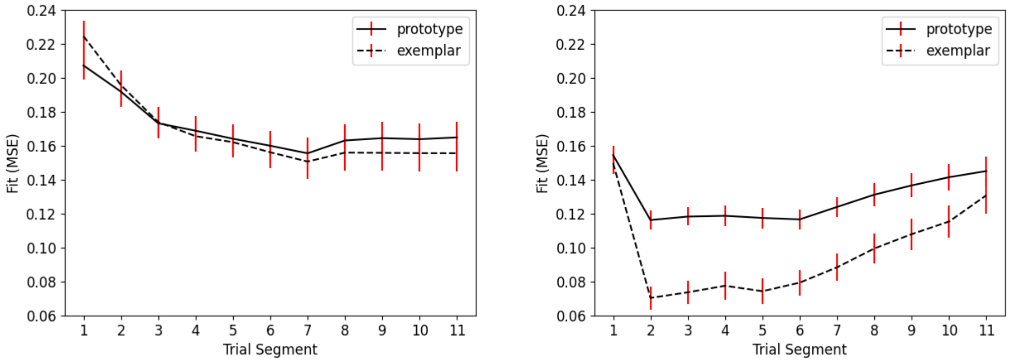
The model results over trial segments from the control condition are compared with the results from Smith & Minda (1998) in Figure 4. The model results over trial segments



(a) Replication

(b) Original

Figure 4: Comparison of (a) the results from our approximate replication in the control condition with (b) the results from the original experiment, reproduced from Smith & Minda (1998). Models were fit for each participant's performance over a trial segment and then averaged over all participants for the condition, with fit calculated as the SSE in a trial segment. Error bars denote the standard error of the mean.



(a) Control Condition

(b) Experimental Condition

Figure 5: Comparison of the prototype and exemplar models in the (a) control and (b) experimental conditions. Models were fit for each participant's performance over the entire experiment and then averaged over all participants for the condition, with fit calculated as the MSE of all per-trial residuals in a trial segment. Error bars denote the standard error of the mean. While the MSE fit was required for the experimental condition due to trial segments being non-contiguous groupings of trials, it was also done for the control condition for ease of comparison.

from the experimental condition are compared with the results from the control condition in Figure 5. As shown in Figures 4 and 5, the advantage of the exemplar model over the prototype model is increasing in the control condition but decreasing in the experimental condition.

To formally test if the difference in model fits between the prototype and exemplar models changed over the course of the experiment, we ran a two-way ANOVA on the model fits with the trial segment (first/second, last) and model type (prototype, exemplar) as within-subject variables. For the analysis of the control condition using the fit measure and model-fitting method derived from Smith & Minda's (1998) experiment, the interaction between model type and trial segment was still significant, $F(1,59) = 19.72$, $p < .001$. The interaction between model type and trial segment was also significant when using our modified model-fitting procedure and calculation of fit for both the experimental condition, $F(1,59) = 80.75$, $p < .001$, and the control condition, $F(1,59) = 21.55$, $p < .001$.

General Discussion

How well do the prototype and exemplar models of categorization reflect human performance over different stages of learning under realistic environmental statistics? In this study, we replicated the experiment by Smith & Minda (1998) in our control condition, in which the frequency of stimulus presentation remains fixed regardless of the last occurrence of the stimulus. We modified this design to present stimuli according to the power-law model of memory decay in order to better reflect realistic environmental statistics (Anderson & Schooler, 1991) in our experimental condition. We hypothesized that setting up the categorization experiment with realistic environmental statistics would reduce the accessibility of exemplar-based representations in the memory over time, therefore reducing or inverting the trend in Smith & Minda's (1998) findings. Our findings from the control condition mostly align with Smith & Minda's (1998) findings that when stimuli are presented with consistent frequency, the exemplar model's advantage over the prototype model grows over time. The results from the experimental condition confirm our hypothesis: when stimulus presentation more accurately reflects the power-law property of the environment, the exemplar model's advantage over the prototype model generally wanes over time.

Average categorization accuracy improved over time in the control condition; this pattern is expected, as participants were shown stimuli with equal frequency throughout the experiment and thus could learn from more information over time. In the experimental condition, however, categorization accuracy steadily declined after the second trial segment. (The first trial segment represents the initial stage of learning, during which accuracy is still expected to be lower.) Therefore, our experimental design was successful at introducing memory constraints that inhibited performance over time. Since later trial segments contain trials further along each object's memory decay curve, when the object is presented less frequently, the behavioral results follow the implications of the recency effect on memory—that individuals have more difficulty retaining objects that were not presented as recently (Anderson & Schooler, 1991).

Our findings help to reconcile incongruities between the results from category learning and memory experiments. Indeed, unlike traditional experimental scenarios involving constant environmental recency functions (Smith & Minda, 1998), memory constraints encountered in realistic environments reduce the advantage of the exemplar model over the prototype model over time. Under such memory constraints, the faster decay of individual exemplars in memory relative to the decay of the prototype (Posner & Keele, 1970; Zeng et al., 2021) more strongly inhibits the performance of the exemplar model over the prototype model later in learning.

Acknowledgements

This work was supported by a C.V. Starr Fellowship awarded to Q.Z.

References

Anderson, J. R., & Schooler, L. J. (1991). Reflections of the environment in memory. *Psychological Science*, *2*, 396–408.

McKinley, S. C., & Nosofsky, R. M. (1995). Investigations of exemplar and decision bound models in large, ill-defined category structures. *Journal of Experimental Psychology: Human Perception and Performance*, *21*, 128–148.

McKinley, S. C., & Nosofsky, R. M. (1996). Selective attention and the formation of linear decision boundaries. *Journal of Experimental Psychology: Human Perception and Performance*, *22*, 294–317.

Medin, D. L. (1975). A theory of context in discrimination learning. In G. H. Bower (Ed.), *The psychology of learning and motivation.* Academic Press.

Medin, D. L., & Schaffer, M. M. (1978). Context theory of classification learning. *Psychological Review*, *85*, 207–238.

Medin, D. L., & Smith, E. E. (1981). Strategies and classification learning. *Journal of Experimental Psychology: Human Learning and Memory*, *7*, 241–253.

Myung, J. I., Pitt, M. A., & Navarro, D. J. (2007). Does response scaling cause the generalized context model to mimic a prototype model? *Psychonomic Bulletin & Review*, *14*, 1043—1050.

Nosofsky, R. M. (1984). Choice, similarity, and the context theory of classification. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *10*, 104–114.

Nosofsky, R. M. (1986). Attention, similarity, and the identification-categorization relationship. *Journal of Experimental Psychology: General*, *115*, 39–57.

Nosofsky, R. M. (1987). Attention and learning processes in the identification and categorization of integral stimuli. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *13*, 87–109.

Nosofsky, R. M. (1988). Exemplar-based accounts of relations between classification, recognition, and typicality. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *14*, 700–708.

Nosofsky, R. M., Palmeri, T. J., & McKinley, S. C. (1994). Rule-plus-exception model of classification learning. *Psychological Review*, *101*, 53–79.

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*, 924–940.

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

Posner, M. I., & Keele, S. W. (1968). On the genesis of abstract ideas. *Journal of Experimental Psychology*, *77*, 353–363.

Posner, M. I., & Keele, S. W. (1970). Retention of abstract ideas. *Journal of Experimental Psychology*, *83*, 304–308.

Reed, S. K. (1972). Pattern recognition and categorization. *Cognitive Psychology*, *3*, 382–407.

Schooler, L. J., & Anderson, J. R. (1997). The role of process in the rational analysis of memory. *Cognitive Psychology*, *32*, 219–250.

Shin, H. J., & Nosofsky, R. M. (1992). Similarity-scaling studies of dot-pattern classification and recognition. *Journal of Experimental Psychology: General*, *121*, 278–304.

Smith, J. D., & Minda, J. P. (1998). Prototypes in the mist: The early epochs of category learning. *Journal of Experimental Psychology*, *24*, 1411–1436.

Smith, J. D., & Minda, J. P. (2002). Distinguishing prototype-based and exemplar-based processes in dot-pattern category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *28*, 800–811.

Smith, J. D., & Minda, J. P. (2011). Prototype models of categorization: Basic formulation, predictions, and limitations. In E. M. Pothos & A. J. Wills (Eds.), *Formal approaches in categorization.* Cambridge University Press.

Zeng, T., Tompary, A., Schapiro, A. C., & Thompson-Schill, S. L. (2021). Tracking the relation between gist and item memory over the course of long-term memory consolidation. *bioRxiv*, 304–308. Retrieved from `https://www.biorxiv.org/content/10.1101/2021.01.05.425378v1` (preprint)