Differentiating Exceptions in Rule-Plus-Exception Category Learning

Yongzhen Xie (yongzhen.xie@mail.utoronto.ca)
Department of Psychology, 100 St. George Street
Toronto, ON M5S 3G3

Michael L. Mack (michael.mack@utoronto.ca)
Department of Psychology, 100 St. George Street
Toronto, ON M5S 3G3

Abstract

The learning of rule-plus-exception categories relies on pattern integration and differentiation, but how the representations of rule-followers and exceptions develop through these two operations remains obscure. Here, we inspected the representational shifts in rule-plus-exception category learning by fitting a computational model to behavioral categorization data. We found that exceptions were differentiated from rule-followers within and between categories through learning. The distanced rule-follower and exception representations in each category formed distinct clusters that together constituted a hierarchically structured categorical representation. Moreover, exception learning increased the representational overlap between rule-followers of opposite categories, thereby blurring the category boundary. Our findings illuminate the representational dynamic underlying the acquisition of rule-plus-exception categories and highlight the roles of pattern integration and differentiation in category learning.

Keywords: category learning; computational modeling; pattern integration; pattern differentiation

Introduction

Nature is full of complex categories that encompass diverse objects. For example, members of the mammal category commonly have four limbs and live on land, but whales, as the category exceptions, have fins and live in the water. Also, whales are confusable with members of the fish category due to their similar appearances and habitats. Yet, people can discriminate whales from fish and classify them under the seemingly dissimilar mammal category. To acquire rule-plus-exception categories, the brain performs pattern integration and differentiation on the category members. Specifically, *pattern integration* increases the overlap of stimulus representations (Brunec et al., 2020; Schlichting & Preston, 2015), whereas *pattern differentiation* reduces the representational overlap (Brunec et al., 2020; Hulbert & Norman, 2015). How do the representations of category rule-followers and exceptions transform through these two operations? We aimed to discern the representational shifts in rule-plus-exception category learning with computational modeling and a novel category learning paradigm.

People can flexibly transform their stimulus representations through category learning. Past behavioral (Goldstone et al., 2001; Juárez et al., 2019; Pothos & Reppa, 2014) and neuroimaging (Dandolo & Schwabe, 2018; Mack et al., 2016) studies indicate that the learning of categories without exceptions can drive within-category stimuli to integrate and between-category stimuli to differentiate. However, such representational shifts can be distorted by the introduction of exceptions. Prior works imply that the brain reduces the representational overlap between rule-followers and exceptions within and between categories (Davis et al., 2012; Heffernan et al., 2021; Sakamoto & Love, 2006). For example, Davis and colleagues (2012) found that a categorization model that differentiated exceptions from rule-followers could predict the activation in the medial temporal lobe during category learning, indicating that the differentiation occurs in this brain region. Other fMRI works (Hulbert & Norman, 2015; Kim et al., 2017) suggest that the hippocampus can differentiate representations of similar but competing events, such as similar-looking exceptions and rule-followers from opposite categories. The differentiation may enable people to identify exceptions as inconsistent members of a category and avoid confusing them with resembling items from a competing category.

Exception learning may also hinder the within-category integration and between-category differentiation of rule-followers indicated by past studies (Goldstone et al., 2001; Pothos & Reppa, 2014). Specifically, Silliman and colleagues (2020) suggest that the presence of exceptions prevents the integration of rule-followers within categories because of the inconsistency between category members. Moreover, rule-follower representations between categories may overlap if they are confusable with exceptions from the competing category (Heffernan et al., 2021). These exception-induced representational changes can blur the category boundary and increase the difficulty of classifying rule-followers.

The existing literature indicates selective pattern integration and differentiation underlying rule-plus-exception category learning, but no one has directly characterized these operations in the learning process. In particular, pattern integration and differentiation involve changes in representational similarities over time (Hulbert & Norman, 2015), but past studies on rule-plus-exception categories often focused on the final learning outcomes (e.g., Heffernan et al., 2021; Sakamoto & Love, 2006). To clarify the integration and differentiation processes during learning, we employed a *delayed exception sequenc*e created by Heffernan and colleagues (2021). They showed that delayed introduction of exceptions in the category learning phase resulted in more precise categorical representations in a hippocampal model in comparison to early introduction of

In J. Culbertson, A. Perfors, H. Rabagliati & V. Ramenzoni (Eds.), *Proceedings of the 44th Annual Conference of the Cognitive Science Society.*

exceptions. Thus, the delayed exception sequence can not only promote the formation of accurate stimulus representations but also enable comparison between representations before and after exception learning. Such comparison would reveal shifts in stimuli's representational similarities induced by the exceptions.

To assess people's latent stimulus representations, we leveraged the Supervised and Unsupervised Stratified Adaptive Incremental Network (SUSTAIN), a model that simulates human categorization (Love et al., 2004). SUSTAIN can develop clusters representative of category members and activate the clusters during the classification of a stimulus. Clusters that better describe the stimulus have higher activations, and the most activated cluster governs the categorization decision. By varying the pattern of cluster activations across stimuli, SUSTAIN can simulate pattern integration and differentiation in the human brain. Neuroimaging studies have shown that SUSTAIN could predict neural operations and representations within various brain regions (e.g., medial temporal cortex, Davis et al., 2012; Mack et al., 2016, 2018; ventromedial prefrontal cortex, Mack et al., 2020; occipitotemporal cortex, Braunlich & Love, 2019). Accordingly, we could fit SUSTAIN to human categorization performance to infer how people represent stimuli latently during category learning.

We tested two predictions to evaluate the representational shifts in rule-plus-exception category learning: (1) Exception learning would result in within- and between-category differentiation between rule-followers and exceptions, and (2) exception learning would hinder within-category integration and between-category differentiation of rule-followers. We fitted SUSTAIN to human categorization data before and after exception learning and performed representational similarity analysis (RSA) on stimuli's cluster activations. Importantly, we observed that exception learning resulted in differentiated clusters of rule-followers and exceptions, in addition to a faded category boundary.

Methods

Participants

We recruited 42 undergraduate students (M_{age} = 18.83, SD_{age} = 2.33; 36 females; 36 right-handed) from the University of Toronto Psychology Participant Pool. Participants completed the study online to receive course credits. Participants gave consent before participation and received debriefing after completing the study.

Stimuli

We created opposite family-resemblance categories (Shepard et al., 1961), in which the stimuli had seven binary feature dimensions. The first six dimensions defined the categories, and the prototypes of the two categories had distinct feature values on these dimensions. Rule-followers in each category shared four of the six defining features of their prototype. Exceptions shared five of the six features of the prototype in the opposite category, making them confusable with non-exception stimuli in the competing category. However, the value of the seventh feature dimension in exceptions differed from the value in non-exception stimuli. Thus, the seventh dimension was irrelevant to the category membership but increased the distinctiveness of exceptions. Each category had a prototype, eight rule-followers, and four exceptions. The prototype, four rule-followers, and two exceptions from each category were shown during learning. The remaining stimuli were shown during testing. All the feature values, except the values of the seventh dimension, appeared equally frequently in the learning and testing phases.

We generated artificial animals (400×500 px) in Photoshop (Figure 1). The animals had seven components, each with two variations. The first six components – including the horn, beak, wings, hand, foot, and tail – were category-defining, and their variations involved simultaneous changes in color and shape. We made these two features covary because people experienced difficulties detecting changes in only the shape in a pilot study. The seventh component, which was the body, tagged exceptions and varied only in color so that it was less salient than the other six components. For each participant, the six category-defining components were randomly matched to the six category-defining feature dimensions. For example, the stimuli learned by one person might have dimensions 1–6 corresponding to horn, beak, wings, hand, foot, and tail, and such correspondences would change for another individual. In this way, participants could learn different stimulus sets with the same category structure.

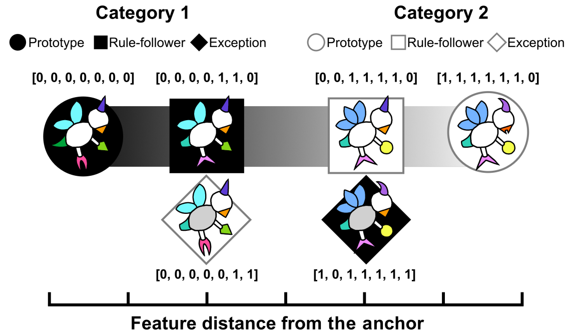

Figure 1: Subway plot of example stimuli and their values on the seven feature dimensions.

Procedure

Our study was approved by the University of Toronto Research Ethics Board. Before the learning phase, we told participants that they would learn two novel animal categories, cordia and naptha. In each trial of the learning phase, an artificial animal appeared for 2.5 s following a 0.7-s fixation cross. Participants then had 3 s to classify the stimulus by pressing a key. We asked them to press "C" on the keyboard if they thought the stimulus was a cordia and "N" if they thought it was a naptha. After responding, participants saw 2-s feedback that included the classified animal and the correctness of their response.

To examine the representational shifts, we introduced prototypes, rule-followers, and exceptions successively in the learning phase (Table 1). Specifically, we split the learning phase into halves, each with four blocks. In the first half, participants learned the prototype and rule-followers from each category in succession. In the second half, participants were introduced to exceptions in each category one at a time while continually learning to classify the prototypes and rule-followers. In the last two blocks, all the learning stimuli were present. Overall, each learning block contained 48 trials, in which the stimuli were repeatedly displayed in random order.

After the first half of the learning phase, participants went through an intermediate testing phase in which they categorized prototypes as well as learned and novel rule-followers without feedback. Exceptions were absent in this phase because their distinct value on the seventh feature dimension could lead participants to guess new category rules. The final testing phase followed the second half of the learning phase and involved all the learned stimuli, "novel" rule-followers from the intermediate testing phase, and novel exceptions.

Table 1: The sequence of stimulus introduction for each category in the learning phase. For rule-followers and exceptions, the numbers inside the brackets indicate individual stimuli. The number after the asterisk indicates how many times individual stimuli were displayed.

Block	Prototype	Rule-followers	Exceptions
1	P * 6	R[1, 2] * 9	
2	P * 6	R[3, 4] * 9	
3	P * 5	R[1, 2, 3, 4] * 4	
4	P * 5	R[1, 2, 3, 4] * 4	
5	P * 8	R[1, 2, 3, 4] * 2	E[1] * 8
6	P * 8	R[1, 2, 3, 4] * 2	E[2] * 8
7	P * 8	R[1, 2, 3, 4] * 2	E[1, 2] * 4
8	P * 8	R[1, 2, 3, 4] * 2	E[1, 2] * 4

Modeling Analysis

Model Fitting We used the CatLearn package (Wills et al., 2017; Wills & Pothos, 2012) in R 4.1.1 to perform the model fitting. The fitting procedure was adapted from the study by Mack and colleagues (2016). We first trained SUSTAIN with trials from the first half of the learning phase using supervised learning. Then, we used *DEoptim* (Mullen et al., 2011), a global optimization algorithm, to optimize the model predictions of participants' categorization accuracies in learning blocks 1–4 and the intermediate testing phase by maximal log-likelihood. Subsequently, we extracted the cluster activations for all the stimuli (i.e., the prototype, eight rule-followers, and four exceptions from each category) before exception learning. With the feature attention weights and clusters developed in the first half of the learning phase, SUSTAIN was trained with trials from the second half of the learning phase and fitted to participants' categorization accuracies in learning blocks 5–8 and the final testing phase. From this re-fitted model, we obtained the cluster activations for stimuli after exception learning.

Representational Similarity Analysis We computed Fisher's *z*-transformed Pearson correlations between stimuli's cluster activations and constructed representational similarity matrices (RSMs) for the intermediate and final testing phases. Based on past studies on pattern integration and differentiation (Brunec et al., 2020; Hulbert & Norman, 2015), a higher correlation between two stimuli reflects a higher overlap of their representations. Accordingly, we quantified the *within-category similarity* (WCS) by averaging the *z*-transformed correlations within categories and the *between-category similarity* (BCS) by averaging the correlations between categories. To evaluate our predictions, we analyzed the within- and between-category similarities between rule-followers and exceptions (i.e., WCS_{RE} and BCS_{RE}) and the within- and between-category similarities of only the rule-followers (i.e., WCS_{R} and BCS_{R}). Because we included multiple exceptions in each category, we also explored the within- and between-category similarities of only the exceptions (i.e., WCS_{E} and BCS_{E}).

Results

Behavioral Results The average categorization accuracies in the learning and testing phases are shown in Figure 2a. We used multivariate analysis of variance (MANOVA) to test the effects of learning blocks on the categorization accuracies for prototypes, rule-followers, and exceptions. For the first half of the learning phase, we found a significant positive main effect of learning blocks on the accuracies for prototypes ($F(3, 164) = 4.98, p = .002, R^2 = .08$) and rule-followers ($F(3, 164) = 3.85$, $p = .01$, $R^2 = .07$), suggesting that individuals' learning performance improved over the first four blocks. The last four blocks did not exert any significant main effect on the accuracies for prototypes ($F(3, 164) = 0.22, p = .88, R^2 = .004$), rule-followers ($F(3, 164) = 0.61, p = .61, R^2 = .01$), and exceptions (E: $F(3, 164) = 1.33, p = .27, R^2 = .02$). Thus, the introduction of exceptions might have impeded people's ability to improve their learning performance.

The testing performance is shown in Figure 2b. We examined participants' learning success by running one-sample t-tests to check if their categorization accuracies in the testing phases were greater than chance (> .5). In the intermediate testing phase, participants achieved significantly above-chance accuracies for prototypes ($M = .96, SE = 0.02, t(41) = 27.18, p < .001$), learned rule-followers ($M = 0.79$, $SE = 0.02$, $t(41) = 12$, $p < .001$), and novel rule-followers ($M = .59$, $SE = 0.03$, $t(41) = 3.36, p < .001$). In the final testing phase, the accuracies were significantly above chance for prototypes ($M = .95$, $SE = 0.02$, $t(41) = 23.2$, $p < .001$), learned rule-followers ($M = .79$, $SE = 0.02$, $t(41) = 14.37$, $p < .001$), novel rule-followers ($M = .59$, $SE = 0.03$, $t(41) = 3.19, p = .001$), and learned exceptions ($M = .7, SE = 0.05$, $t(41) = 4.06$, $p < .001$). The above-chance categorization accuracy for novel rule-followers indicates that participants could generalize the general category rules to unfamiliar items. However, the accuracy for novel exceptions was at the chance level ($M = .53$, $SE = 0.05$, $t(41)$

= 0.67, *p* = .25). This result suggests that individuals experienced difficulties generalizing patterns of exceptions.

We assessed the link between intermediate and final testing performance. We found a significant positive correlation between the categorization accuracies in the intermediate and final testing phases for non-exception stimuli (prototypes and rule-followers; *r* = .48, *t*(40) = 3.46, *p* = .001). The intermediate testing accuracy for non-exception stimuli was not significantly correlated to the final testing accuracy for exceptions (*r* = -.07, *t*(40) = -0.48, *p* = 0.64). We compared the two correlations using Pearson and Filon's *z* in the cocor package (Diedenhofen & Musch, 2015). We found that the correlations were significantly different (*z* = 3.27, *p* = .001). Thus, participants classified non-exception stimuli more accurately in the final testing phase if they performed better on these items before exception learning. In contrast, the intermediate testing performance on non-exception stimuli was not predictive of how participants categorized exceptions in the final testing phase.

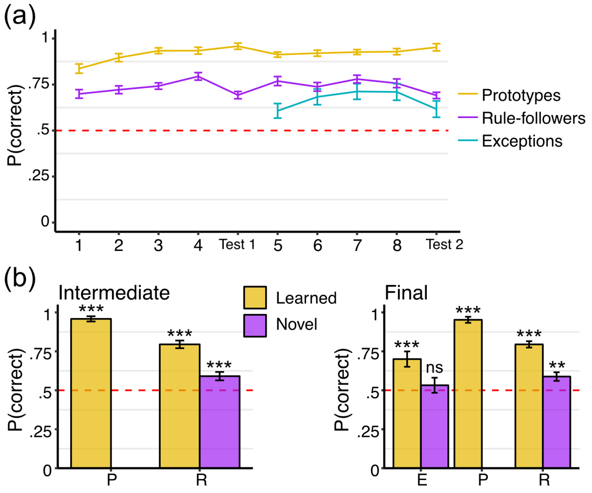

Figure 2: (a) Human categorization accuracies in learning blocks 1–8 and testing phases. Test 1: Intermediate testing; Test 2: Final testing. (b) Human categorization accuracies for learned and novel stimuli in testing phases. The asterisks represent the statistical significance of one-sample t-tests. ****p* < .001; ***p* < .01; ns: *p* > .05. Error bars: ± *SE*. P: Prototypes; R: Rule-followers; E: Exceptions.

Model Fitting Results We evaluated SUSTAIN's testing performance because we aimed to infer people's stimulus representations in the testing phases. The model testing performance is shown in Figure 3. The model's categorization accuracies in the testing phases were significantly correlated to participants' accuracies for all stimulus types except the prototypes in the final testing phase (Table 2). This non-significant correlation could be due to the model and participants' near-ceiling-level performance on the prototypes. Moreover, we found a significant positive correlation between the model's intermediate and final testing accuracies for non-exception stimuli (*r* = .54, *t*(40) = 4.08, *p* < .001). However, the intermediate testing accuracy for non-exception stimuli was not significantly correlated to the final testing accuracy for exceptions (*r* = .18, *t*(40) = 1.15, *p* = .26). Pearson and Filon's *z* test from the cocor package (Diedenhofen & Musch, 2015) revealed a significant difference between the two correlations (*z* = 2.49, *p* = .01). Therefore, like human participants, the model's intermediate testing performance predicted its final testing performance on non-exception stimuli but not exceptions.

Table 2: Correlations between model and human categorization accuracies in the testing phases. The asterisks indicate the statistical significance of each correlation. ****p* < .001; ***p* < .01; ns: *p* > .05.

	Prototype	Rule-followers	Exceptions
	Intermediate Testing Phase		
Learned	.72***	.8***	
Novel		.81***	
	Final Testing Phase		
Learned	.01(ns)	.68***	.77***
Novel		.83***	.41**

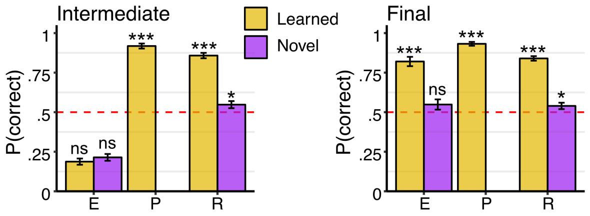

Figure 3: Model categorization accuracies for learned and novel stimuli in testing phases. The asterisks represent the statistical significance of one-sample t-tests that checked if the categorization accuracies were above chance (> .5). ****p* < .001; **p* < .05; ns: *p* > .05. Error bars: ± *SE*. P: Prototypes; R: Rule-followers; E: Exceptions.

RSA Results The RSMs for the intermediate and final testing phases (Figure 4a) revealed shifts in stimulus representations through exception learning. We performed two-tailed repeated-measures t-tests to assess how the similarity scores (i.e., WCS and BCS) for rule-followers and exceptions changed between testing phases (Figure 4b). We found that WCS$_{\mathrm{RE}}$ decreased significantly in the final testing phase (intermediate: *M* = -0.48, *SE* = 0.09; final: *M* = -0.76, *SE* = 0.07; final vs. intermediate: *t*(41) = -3.14, *p* = .003), suggesting a reduction in the representational overlap between within-category rule-followers and exceptions during exception learning. In addition, BCS$_{\mathrm{RE}}$ showed a significant decrease (intermediate: *M* = 0.9, *SE* = 0.11; final: *M* = -0.23, *SE* = 0.07; final vs. intermediate: *t*(41) = -7.14, *p* < .001), indicating that the representations of between-category rule-followers and exceptions became less overlapped through exception learning. These results support our prediction that rule-followers and exceptions would undergo pattern differentiation within and between categories in the learning process.

We assessed changes in the representational similarities of only the rule-followers. We found a significant increase in $\mathrm{WCS_R}$ across the intermediate and final testing phases (intermediate: $M = 0.92$, $SE = 0.1$; final: $M = 1.49$, $SE = 0.09$; final vs. intermediate: $t(41) = 5.15$, $p < .001$), which indicates the integration of within-category rule-followers. This result challenges our prediction that exception learning would disrupt the overlapping of rule-followers from the same category. Furthermore, $\mathrm{BCS_R}$ increased significantly across the testing phases (intermediate: $M = -0.15$, $SE = 0.1$; final: $M = 0.83$, $SE = 0.11$; final vs. intermediate: $t(41) = 6.29$, $p < .001$), implying that the representations of between-category rule-followers were integrated. The result is consistent with our prediction that exception learning would hamper the between-category differentiation of rule-followers.

We explored shifts in the representational similarities of only the exceptions. $\mathrm{WCS_E}$ showed a significant increase in the final testing phase (intermediate: $M = 2.09$, $SE = 0.09$; final: $M = 2.36$, $SE = 0.07$; final vs. intermediate: $t(41) = 2.72$, $p = .01$), which suggests that exceptions from the same category were integrated during exception learning. Similarly, $\mathrm{BCS_E}$ increased significantly (intermediate: $M = -0.89$, $SE = 0.1$; final: $M = 0.31$, $SE = 0.12$; final vs. intermediate: $t(41) = 7.63$, $p < .001$), indicating that exception representations from competing categories became overlapped. Therefore, like the rule-follower representations, exception representations underwent pattern integration within and between categories through learning.

Multidimensional Scaling of RSMs Our RSA results indicate that exception learning not only causes differentiation between rule-followers and exceptions but also drives these two types of stimuli to undergo separate integration. Consequently, specialized clusters for rule-followers and exceptions may form in the representational space. To visualize the clustering of stimulus representations, we converted each RSM into a distance matrix by computing the absolute difference between each z-transformed correlation and the z-transformed correlation between the cluster activations of the same stimulus (i.e., the maximum correlation in the RSM). Thus, a higher value in the distance matrices reflected higher dissimilarity between two stimuli's activation patterns. Then, we ran metric multidimensional scaling (MDS; $k = 2$) on the distance matrices to visualize stimuli's distributions on a two-dimensional plane.

The stimulus distributions in the intermediate and final testing phases are shown in Figure 4c. Before exception learning, exception representations were intermixed with rule-follower representations from the competing category due to the perceptual similarity. After learning, rule-follower and exception representations in each category formed their distinct, non-overlapping clusters. Also, exception learning led rule-follower representations from competing categories to overlap. Altogether, these changes in stimulus distributions suggest that exception learning results in the formation of differentiated rule-follower and exception clusters and the integration of between-category rule-followers in the representational space.

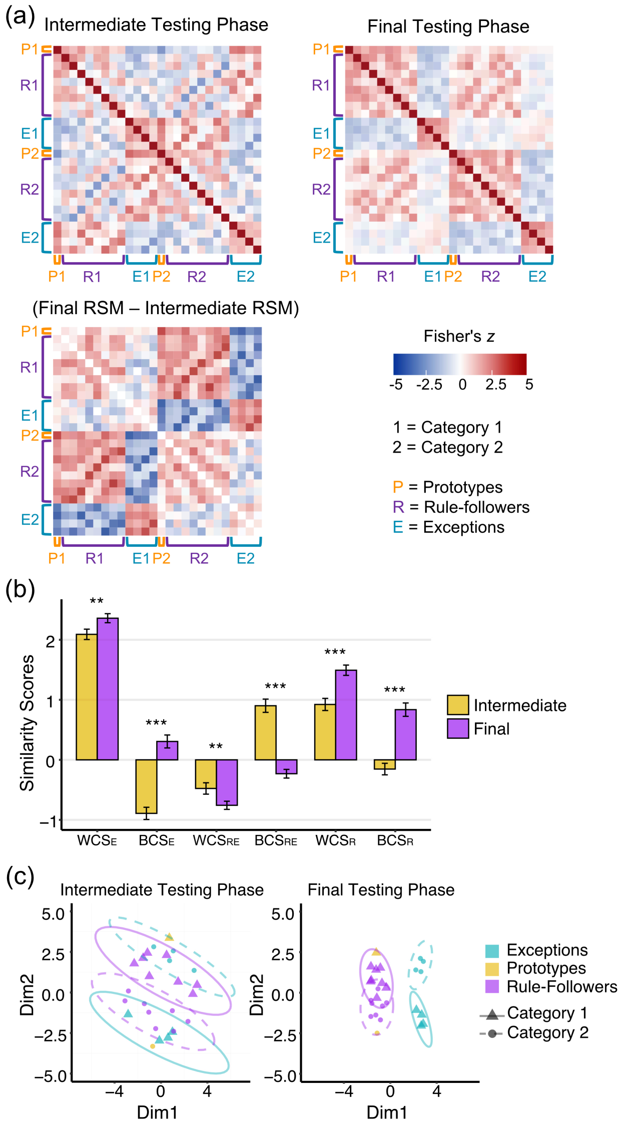

Figure 4: (a) RSMs in testing phases. Each cell represents a z-transformed correlation between the cluster activations of two stimuli. (b) Representational similarity scores in testing phases. The asterisks indicate the statistical significance of repeated-measures t-tests. $^{***}p < .001$, $^{**}p < .01$. Error bars: ± SE. (c) Distributions of stimuli's cluster activation patterns projected onto a two-dimensional MDS space. The ellipses separately group rule-followers and exceptions in each category based on the multivariate t-distribution.

Discussion

We investigated transformations of stimulus representations during rule-plus-exception category learning. Matching our predictions, we found that exception learning drove pattern differentiation between rule-followers and exceptions within and between categories. Also, the learning led to the pattern integration of between-category rule-followers. Contrary to our prediction, within-category integration of rule-followers happened during exception learning. We further showed that rule-followers and exceptions in each category formed their unique representational clusters. In summary, our findings help discern the detailed representational shifts underlying the learning of rule-plus-exception categories.

We showed that exception learning resulted in within- and between-category differentiation between rule-followers and exceptions. The dissociated rule-follower and exception representations might explain why people's intermediate testing accuracies for non-exception stimuli predicted the final testing accuracies for only the non-exception but not the exception stimuli. Our results align with findings from past neuroimaging (Davis et al., 2012), behavioral (Sakamoto & Love, 2006), and neural modeling (Heffernan et al., 2021) studies that indicate that the human brain constructs distinct exception representations during rule-plus-exception category learning. However, none of the past studies directly characterized the differentiation process that leads to those distinct representations. Here, we leveraged a computational model and a delayed exception sequence to fill the gap in the existing literature, providing novel evidence of selective pattern differentiation in the category learning process.

The differentiation between rule-follower and exception representations may support the identification of exceptions, especially when these items are confusable with members of the competing category. The pattern differentiation between confusable events has been found within the hippocampus in prior fMRI works (Hulbert & Norman, 2015; Kim et al., 2017). Such works hint that the perceptual similarity between rule-followers and exceptions from competing categories may govern the between-category differentiation between these two types of stimuli. In contrast, the within-category differentiation between rule-followers and exceptions may be driven by their inconsistent patterns. Future works can delve into the determinants of the within- and between-category differentiation and the specific contributions of these operations to category learning.

We found that within-category integration of rule-followers occurred during exception learning, contrary to our prediction that the learning would hinder this operation. Our result also contradicts Silliman and colleagues (2020)' s findings: They showed that exception learning prevented the typical increase in the perceived similarity of non-exception stimuli within categories. The inconsistent findings may be due to differences in the learning sequences the two studies used. We introduced exceptions in only the second half of the learning phase, whereas Silliman and colleagues (2020) introduced all the stimuli at the beginning of the learning phase. As implied by Heffernan and colleagues (2021), the delayed introduction of exceptions, compared to the early introduction, can result in more established similarity representations of within-category rule-followers that are less susceptible to the distortions induced by subsequent exception learning. Overall, the opposite findings from the present and past studies necessitate a comparison between the representational shifts in early and delayed exception learning conditions.

We observed that rule-followers between categories underwent pattern integration during exception learning. Particularly, the inconsistency in members of the same category and the overlap of members from competing categories may drive the between-category integration and blur the category boundary. Indeed, similar boundary-blurring during the learning of rule-plus-exception categories was observed in the modeling study by Heffernan and colleagues (2021). The overlapping of between-category rule-followers potentially hinders category learning and explains our finding that people's learning performance failed to improve after the introduction of exceptions.

We revealed that exceptions between categories became integrated through exception learning, which further hints at the blurred category boundary. Moreover, we found pattern integration of within-category exceptions during learning. The separate integration of rule-followers and exceptions, combined with the differentiation between these stimulus types, gave rise to distinct rule-follower and exception clusters in the representational space. Our results are congruent with the behavioral findings by Savic and Sloutsky (2019), which imply that people represent within-category exceptions as a subgroup away from the rule-followers in the same category. The rule-follower and exception clusters within a category can serve as subcategories that constitute a hierarchically structured categorical representation. In support of the notion of the hierarchical structure, past works have shown that the human brain and artificial neural networks could divide within-category stimuli into subgroups based on their perceptual dissimilarities (Ahn et al., 2021; Konkle & Alvarez, 2022). The hierarchical representational structure may allow people to distinguish patterns that characterize different within-category stimuli, such as the distinct patterns for rule-followers and exceptions.

Our study provides a novel modeling approach to inspect the latent representational shifts during category learning. Specifically, we leveraged SUSTAIN to examine people's representations because past works suggest that this model can predict neural activities related to category learning (e.g., Davis et al., 2012; Mack et al., 2016). However, we expect any model that supports pattern integration and differentiation, such as the hippocampal model (Schapiro et al., 2017), to allow inference of the representational shifts. Future works can extend our modeling approach to the acquisition of other types of categories. Future works can also use the representational shifts in the model to predict changes in the neural representations during rule-plus-exception category learning to deepen the understanding of the neural mechanisms underlying the learning process.

In conclusion, we combined a computational model and a delayed exception sequence to shed new light on how stimulus representations transform in rule-plus-exception category learning. By studying category learning at the representational level, we can discern the operations fundamental to this cognitive process and understand the acquisition of sophisticated categories in the real world.

References

Ahn, S., Zelinsky, G. J., & Lupyan, G. (2021). Use of superordinate labels yields more robust and human-like visual representations in convolutional neural networks. *Journal of Vision*, *21*(13), 13.

Braunlich, K., & Love, B. C. (2019). Occipitotemporal representations reflect individual differences in conceptual knowledge. *Journal of Experimental Psychology. General*, *148*(7), 1192–1203.

Brunec, I. K., Robin, J., Olsen, R. K., Moscovitch, M., & Barense, M. D. (2020). Integration and differentiation of hippocampal memory traces. *Neuroscience & Biobehavioral Reviews*, *118*, 196–208.

Dandolo, L. C., & Schwabe, L. (2018). Time-dependent memory transformation along the hippocampal anterior–posterior axis. *Nature Communications*, *9*(1), 1205.

Davis, T., Love, B. C., & Preston, A. R. (2012). Learning the exception to the rule: Model-based FMRI reveals specialized representations for surprising category members. *Cerebral Cortex (New York, N.Y.: 1991)*, *22*(2), 260–273.

Diedenhofen, B., & Musch, J. (2015). cocor: A comprehensive solution for the statistical comparison of correlations. *PLOS ONE*, *10*(4), e0121945.

Goldstone, R. L., Lippa, Y., & Shiffrin, R. M. (2001). Altering object representations through category learning. *Cognition*, *78*(1), 27–43.

Heffernan, E. M., Schlichting, M. L., & Mack, M. L. (2021). Learning exceptions to the rule in human and model via hippocampal encoding. *Scientific Reports*, *11*(1), 21429.

Hulbert, J. C., & Norman, K. A. (2015). Neural differentiation tracks improved recall of competing memories following interleaved study and retrieval practice. *Cerebral Cortex (New York, N.Y.: 1991)*, *25*(10), 3994–4008.

Juárez, F. P.-G., Sicotte, T., Thériault, C., & Harnad, S. (2019). Category learning can alter perception and its neural correlates. *PLOS ONE*, *14*(12), e0226000.

Kim, G., Norman, K. A., & Turk-Browne, N. B. (2017). Neural Differentiation of Incorrectly Predicted Memories. *The Journal of Neuroscience*, *37*(8), 2022–2031.

Konkle, T., & Alvarez, G. A. (2022). A self-supervised domain-general learning framework for human ventral stream representation. *Nature Communications*, *13*(1), 491.

Love, B. C., Medin, D. L., & Gureckis, T. M. (2004). SUSTAIN: A network model of category learning. *Psychological Review*, *111*(2), 309–332.

Mack, M. L., Love, B. C., & Preston, A. R. (2016). Dynamic updating of hippocampal object representations reflects new conceptual knowledge. *Proceedings of the National Academy of Sciences*, *113*(46), 13203–13208.

Mack, M. L., Love, B. C., & Preston, A. R. (2018). Building concepts one episode at a time: The hippocampus and concept formation. *Neuroscience Letters*, *680*, 31–38.

Mack, M. L., Preston, A. R., & Love, B. C. (2020). Ventromedial prefrontal cortex compression during concept learning. *Nature Communications*, *11*(1), 46.

Mullen, K. M., Ardia, D., Gil, D. L., Windover, D., & Cline, J. (2011). DEoptim: An R package for global optimization by differential evolution. *Journal of Statistical Software*,

Pothos, E. M., & Reppa, I. (2014). The fickle nature of similarity change as a result of categorization. *Quarterly Journal of Experimental Psychology*, *67*(12), 2425–2438.

Sakamoto, Y., & Love, B. C. (2006). Vancouver, Toronto, Montreal, Austin: Enhanced oddball memory through differentiation, not isolation. *Psychonomic Bulletin & Review*, *13*(3), 474–479.

Savic, O., & Sloutsky, V. M. (2019). Assimilation of exceptions? Examining representations of regular and exceptional category members across development. *Journal of Experimental Psychology. General*, *148*(6), 1071–1090.

Schapiro, A. C., Turk-Browne, N. B., Botvinick, M. M., & Norman, K. A. (2017). Complementary learning systems within the hippocampus: A neural network modelling approach to reconciling episodic memory with statistical learning. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *372*(1711), 20160049.

Schlichting, M. L., & Preston, A. R. (2015). Memory integration: Neural mechanisms and implications for behavior. *Current Opinion in Behavioral Sciences*, *1*, 1–8.

Shepard, R. N., Hovland, C. I., & Jenkins, H. M. (1961). Learning and memorization of classifications. *Sychological Monographs: General and Applied*, *75*(13), 1–42.

Silliman, D. C., Snoddy, S., Wetzel, M., & Kurtz, K. J. (2020). Costly exceptions: Deviant exemplars reduce category compression. *Proceedings of the 42nd Cognitive Sciences Society Meeting*, 7.

Wills, A. J., O'Connell, G., Edmunds, C. E. R., & Inkster, A. B. (2017). Progress in modeling through distributed collaboration: Concepts, tools and category-learning examples. In B. H. Ross (Ed.), *Psychology of Learning and Motivation* (Vol. 66). Academic Press.

Wills, A. J., & Pothos, E. M. (2012). On the adequacy of current empirical evaluations of formal models of categorization. *Psychological Bulletin*, *138*(1), 102–125.

Acknowledgments

We thank E. Heffernan for the inspiration for the experimental design, G. Son for suggestions on data collection, and M. Schlichting for suggestions on data analysis. Our research is supported by the Natural Sciences and Engineering Research Council (NSERC) Discovery Grant to MLM (RGPIN-2017-06753), the Canada Foundation for Innovation and Ontario Research Fund (36601) to MLM, the Canada Institutes of Health Research (CIHR) Project Grant (PJT-178337) to MLM, the Brain Canada Future Leaders in Canadian Brain Research Grant to MLM, and the Ontario Graduate Scholarship (OGS) to YX.