Labels Facilitate Categorical Perception Effects during Novel Category Learning

Andrew J. Mertens[1] (andrew.mertens@colorado.edu), Eliana Colunga[1,2] (colunga@colorado.edu), & Albert Kim[1,2] (albert.kim@colorado.edu)

[1]Department of Psychology & Neuroscience, University of Colorado Boulder, Boulder, Colorado, USA

[2]Institute of Cognitive Science, University of Colorado Boulder, Boulder, Colorado, USA

Abstract

Previous work has suggested that category labels can facilitate faster rates of category learning. To better understand the role that labels play in this phenomenon, the present study investigates whether this labeling advantage coincides with a warping of representational space that is indicative of categorical perception. To this end, we collected behavioral and EEG data during two tasks: an approach-avoid task in which category boundaries are learned and a same-different task. The behavioral results replicate the labeling advantage in category learning and suggest that CP effects are strengthened by the influence of labels. Representational similarity analyses of EEG brain activity collected during the approach-avoid task provide additional support for this theory, showing that stimulus representations exhibit patterns of representational warping that are characteristic of CP when the labeling advantage is most prominent. Together, these findings contribute to a richer understanding of how labels facilitate category learning.

Keywords: Labeling; Novel Category Learning; Categorical Perception; EEG; RSA

Introduction

Categories allow us to separate things that differ to us in meaningful ways while simultaneously grouping together the ones that share features in common that we find important. The mere act of learning to make these crucial distinctions and form these like groupings has been shown to have a measurable influence on the way we experience the individual items upon which we impose these groupings. In a phenomenon referred to as **categorical perception** (CP), items belonging to the same category begin to appear relatively more similar to one another and items from differing categories begin to appear relatively more distinct from one another (Goldstone & Hendrickson, 2010). CP effects can be conceptualized as the warping of representational space for the items in question, such that the representations of items in the same category move closer together and the representational space between categories grows.

Previous work has suggested that labels facilitate novel category learning despite providing redundant category information (Lupyan et al., 2007; Fotiadis & Protopapas, 2022). This **labeling advantage** consists of increased (i.e., steeper) category learning rates in the label condition compared to the no label condition. Notably, this effect is typically only present relatively early in category learning, with accuracy rates in the no label condition eventually reaching the same level as those of the label condition. This work sought to improve understanding of this intriguing phenomenon by examining whether labels and the advantage they provide influence the formation of CP effects during category learning.

Current Study

Given that category learning, with or without labels, is associated with the formation of CP effects, we aimed to measure the extent to which stimulus representations are warped to reflect CP during category learning and examine how labels influence that warping process. In this exploratory study, participants learned novel category classifications in an approach-avoid task with or without the support of labels. We measured CP effects using a combination of behavioral and electroencephalographic (EEG) measures. Behaviorally, we measure CP effects by comparing accuracy for identifying within- and between- category differences in a same-different task interspersed between blocks of the approach-avoid task. If category learning influences stimulus representations of between-category items to become more distinct and within-category items to become more similar, then accuracy for detecting between-category differences will improve more than accuracy for detecting within-category differences during category learning, even when matched for physical differences. Using representational similarity analysis (RSA), we measured CP effects by evaluating whether patterns of EEG brain activity elicited from the stimuli of the same category were more similar than those of stimuli belonging to different categories. The RSA not only affords us an additional way to evaluate the formation of CP effects on the timescale of category learning overall but also on the timescale of an individual trial, which may allow us to glean some information about the level of perceptual analysis in which labels facilitate category learning.

We hypothesized that labels would facilitate category learning and that, during the course of category learning, labels would be associated with the formation of stronger representational warping, particularly during the portion of category learning when the labeling advantage was most prevalent.

Method

Participants

Forty-three undergraduate students participated in this experiment in exchange for course credit. Two participants were excluded from analysis due to an excessive number of rejected trials (>10%) resulting from EEG artifacts. The remaining

In D. Barner, N.R. Bramley, A. Ruggeri and C.M. Walker (Eds.), *Proceedings of the 47th Annual Conference of the Cognitive Science Society*

41 participants included in the analysis ranged from age 18 to age 26 (M=19.71, SD=1.83). Twenty-nine of these participants self-identified as female and 12 identified as male. Twenty-five participants self-identified as White, seven as Hispanic or Latino, two as American Indian or Alaskan Native, three as Asian, one as Arabic, two as multiracial, and one opted not to self-identify.

Stimuli

The visual stimuli were purple alien figures with a Gabor patch in the face region, which was described as the alien's eye to participants (see Figure 2). All aliens shared the same body but differed in the Gabor patch eye. Gabor patches varied in orientation and spatial frequency, with eight values on each dimension, resulting in 64 possible combinations (Figure 2). The orientations to create this stimuli consisted of four orientations centered around a vertical orientation (-33.75, -11.25, 11.25, and 33.75 degrees, vertical being 0°) and four centered around a horizontal orientation (56.25, 78.75, -78.75, and -56.25 degrees, vertical being 0°). The eight spatial frequencies used to create these stimuli increase by a factor of 1.2 in units of cycles per degree (cpd) of visual angle with 2 cpd being the lowest: 2, 2.24, 2.51, 2.81, 3.15, 3.52, 3.95, and 4.42 cpd.

The Gabor patches subtended approximately 1.4° x 1.4° of visual angle. Nonce labels presented in the label condition consisted of auditory recordings of a male voice saying gowachi (gOh-wOch-ee) and havnori (hAv-nOR-ee).

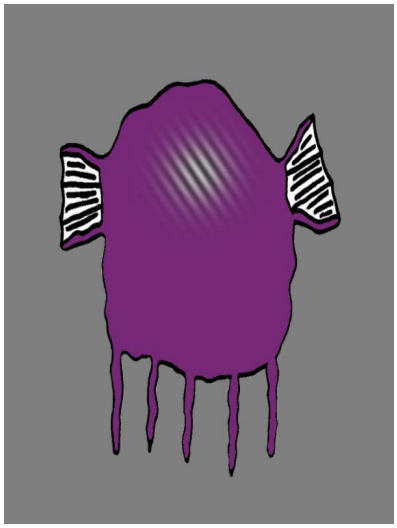

Figure 1: One example of the alien stimuli used in the category learning and same-different judgment tasks.

Design

Participants completed an **approach-avoid** task in which they classified a series of alien stimuli into one of two groups: hostile or friendly aliens. Participants were required to learn to discriminate between these two groups based on physical differences in the orientation dimension. The decision to train

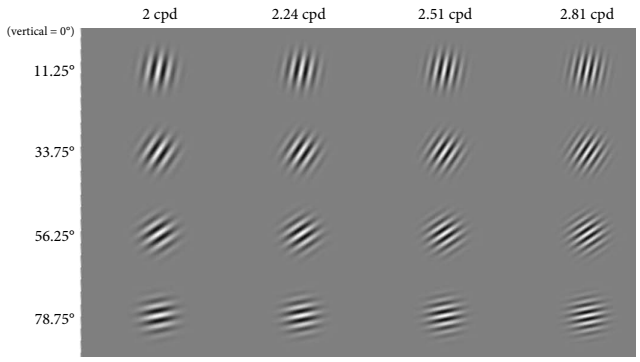

Figure 2: A subset of the 64 unique Gabor patches represented by the combinations of different spatial frequencies and orientations. The two top rows represent stimuli belonging to the friendly category and the lower two represent the hostile category.

only orientation-based categorization is informed by our previous work showing that, while overall accuracy for learning categories based on spatial frequency is higher, the labeling advantage is only apparent in orientation-based categorization (see Mertens & Colunga, 2024 for more discussion). The four relatively vertical orientations corresponded to the friendly classification of alien and the four relatively horizontal orientations represented hostile ones. During category learning, half of the participants solely received feedback indicating whether their classifications were correct or incorrect (no label condition) and the other half heard nonce labels corresponding to the category they saw in addition to the feedback (labeled condition). These labels are redundant in that the classification feedback heard by all participants provides all of the category information needed to learn the task. This task was completed in two parts, each consisting of 97 trials presented in a quasi-random order. The order in which participants saw these two quasi-random sequences was counterbalanced.

Participants also completed a **same-different** judgment task. In each trial of this task, a pair of aliens from the stimulus pool is shown and participants are asked to indicate whether the aliens in each pair were identical to one another or not. Half of the trials displayed a pair of identical aliens, 25% displayed a pair of aliens whose eyes differed in spatial frequency, and 25% displayed a pair of aliens whose eyes differed in orientation. Each 'different' trial consisted of pairs of aliens that differed in either orientation or spatial frequency by 1-3 steps with regard to the stimulus values used in each dimension. For example, a trial in which a participant saw aliens with Gabor patch eyes at orientations of 11.25, and 78.75 degrees would represent three steps of orientation separation (i.e., three increments of 22.5°). Additionally, among the trials in which aliens differed in orientation, half represented between-category differences spanning the category boundary and half represented within-category

differences. This task was completed in three blocks, each consisting of 256 trials.[1] In this task, higher accuracy in the between-category condition compared to the within-category condition is an indication of representational warping. Within each block, each sequence of 16 trials was balanced such that it contained the same number of one, two, and three-step differences in orientation among between-category and within-category trials.

Procedure

The experiment was implemented using Psychopy behavioral testing software (Peirce et al., 2019). The stimuli were viewed on a 1600- by 1024-pixel, 30 by 46.5 cm display from a viewing distance of approximately 136 cm. During the experiment, participants were seated comfortably with the keyboard resting on their lap in a dimly lit and sound-shielded testing room. The two halves of the approach-avoid task were interleaved between the three blocks of the same-different task (see Figure 3). The experiment, including setup procedures, took the average participant one hour to one hour and thirty minutes.

In the same-different task, after being presented with a fixation cross for 500 ms, two alien stimuli were presented, one on each side of the fixation point, for 200 ms before disappearing. Participants were instructed to indicate by key press whether the pair of aliens was identical ('s') or different ('d'), with no time limit for their responses. A blink break occurred after each sequence of four trials.

Within each trial of the approach-avoid task, a fixation cross would be presented for 500 ms followed by an alien stimulus and an illustration of a space explorer. The explorer would appear in one of four locations relative to the alien: above, below, to the left, or to the right. The participant was tasked with classifying each alien as friendly or hostile by pressing the arrow key corresponding to moving the explorer closer to the alien, in the case of a friendly stimulus, or away from the alien, in the case of a hostile. Once the participant made a response, they would hear either a 'bloop' sound in the case of a correct response or a 'buzz' sound in the case of an incorrect response. In the label condition, participants would hear the nonce label associated with the category of alien in that trial. Lastly, we included a break allowing the participant to blink freely, which lasted until the participant opted to proceed to the next trial. The alien and explorer of that trial remained on the screen up to this point.

EEG Recording

Continuous EEG was recorded from an array of 63 cap-embedded active electrodes (BrainVision ActiCap) during all blocks of both tasks. Two additional electrodes monitoring eye movements and blinks, and another two were placed over the left and right mastoid bones. The EEG was amplified, filtered with a DC to 200 Hz bandpass filter, and digitized at 1000 Hz (BrainVision ActiChamp). Electrode impedance was kept below 10 kΩ.

EEG Preprocessing

The offline preprocessing steps described below were performed on an individual participant basis in the MATLAB environment using EEGLAB (Delorme & Makeig, 2004) and ERPLAB (Lopez-Calderon & Luck, 2014) packages. EEG data were downsampled to 250 Hz. Each participant's data was highpass filtered at a 1Hz cutoff and submitted to Independent Components Analysis (ICA) to identify IC's reflecting eye movements and blinks, which were subsequently removed from the unfiltered data. The data were then bandpass filtered (0.5 to 30 Hz) and re-referenced to the average of the mastoid electrodes. We removed further noise from the EEG signal by performing a spatial principle components analysis on all trials and then reconstructing the scalp-EEG data after removing low-ranking components and retaining components that accounted for 95% of the total variance.

An epoch was created for each trial, consisting of a 750 ms segment of the data time-locked to the onset of the display of alien, plus a 200 ms pre-stimulus baseline interval. Trials containing a deflection greater than 100 μV during the epoch were excluded from analysis. EEG data was standardized across trials using z-score transformations to account for individual differences between participants.

Computing Representational Similarity

For our RSA, representational vectors corresponding to each orientation were formed from scalp-wide EEG activity in each trial of the approach-avoid task. First, a window of EEG activity from 60-750 ms post-stimulus onset was taken from each electrode of each trial of the approach-avoid task. Trials with response times of under 500 ms were excluded to ensure that these windows of activity would not contain any neural response to perceiving auditory labels, which are presented 250 ms after a response is made. This criterion resulted in an average of 2.24 excluded trials per participant, with the maximum number of trials excluded for any participant being 27 (i.e., 13.92%). The number of excluded trials did not significantly differ between labeling conditions, (p=.411). We then segmented the 690 ms of trial-level EEG activity recorded at each channel using a sliding window average with a window size of 40 ms and a step size of 20 ms, resulting in 36 time segments per trial.

Within each block, each orientation was presented four times across different trials. For each time segment, we averaged voltage potentials across the timepoints contained in that segment window and across the four trials corresponding to each orientation, then concatenated these averages from all electrode channels, resulting in a representational vector of each orientation at each segment. We then used these vectors to calculate cosine similarity between pairs of orientations at each segment. We compare each pair of stimuli that represents one, two, or three steps of separation between ori-

[1] The first 12 participants we recruited completed a version of the task in which 128 trials were presented per block. The mixed effect models used in this study have been shown to be resilient to unequal numbers of observations per individual (Heo & Leon, 2005)

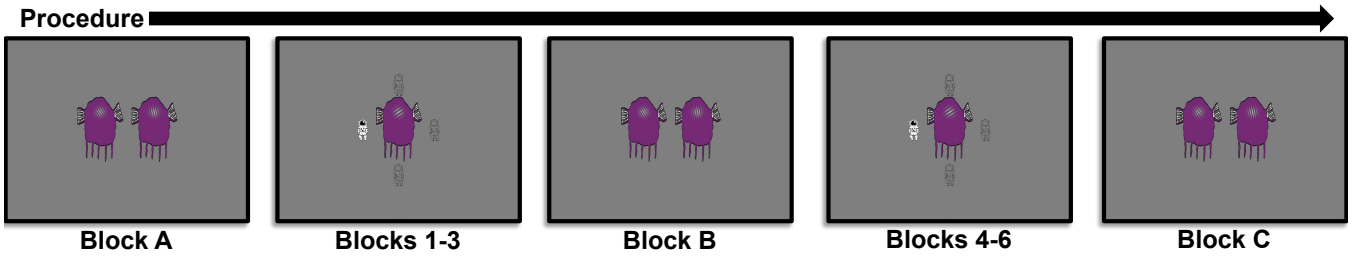

Figure 3: A schematic demonstrating the interleaving of the same-different task (blocks A, B, and C) and the approach-avoid task (blocks 1-6).

entations, resulting in a six-by-six similarity matrix for each segment in which objective perceptual similarity of between-category pairs is balanced with that of within-category pairs. Within each similarity matrix, half of the values represent the representational similarity of pairs of stimuli belonging to the same category (e.g., two friendly alien stimuli with relatively vertically oriented gratings paired together) and half represent similarity of pairs belonging to different categories (e.g., a friendly alien stimulus with a relatively vertically oriented grating paired with a hostile alien stimulus with a relatively horizontally oriented grating). We averaged each of these groups of similarity values together, allowing us to compute separate similarity scores for pairs of orientations that represent within- and between-category differences.

Behavioral Results

Approach-Avoid Task

First, we examined performance in the approach-avoid task to test whether the labeling advantage was present. For the purpose of analysis, the two halves of the approach-avoid task were each divided into 3 blocks of 32 trials (the first trial of each half was excluded as practice). In a mixed effects logistic regression model, single-trial accuracy was fit to fixed effects of block, label condition, and sequence order (i.e., the order in which the two 97-trial sequences that made up each half of the task were presented) in addition to a fixed interaction term for block and label. Random intercepts were included for participants. A likelihood ratio test (LRT) was conducted to evaluate fixed effects. We found that accuracy improved significantly over the course of category learning, as indicated by a main effect of block, $\chi^2(5)=132.38$, $p<.001$. A labeling advantage effect was also revealed, such that accuracy in the label condition increased at a faster rate than in the no label condition. This is indicated by a significant interaction between block and label $\chi^2(5)=17.30$, $p=.004$ (see Figure 4). No main effect of label was revealed ($p=.355$). Although accuracy in the label condition was numerically higher than in the no label condition in all but block three, Tukey pairwise comparisons revealed that differences in accuracy between labeling conditions were not significant in any single block.[2]

[2]We tested an additional model including a fully interacted term for number of trials per block in the same-different task (128 or 256) and verified that this change did not affect category learning.

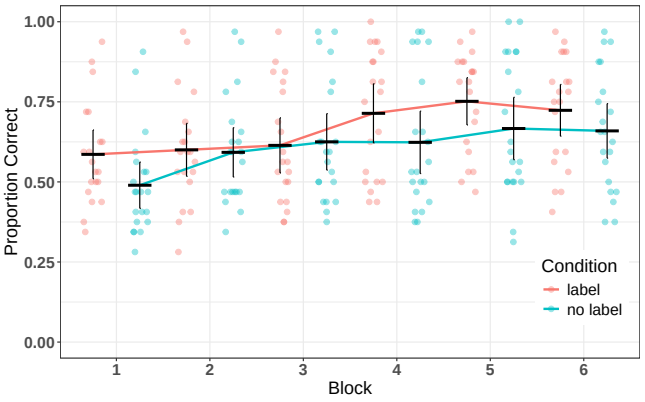

Figure 4: The emergence of the labeling advantage during the approach-avoid task. Horizontal bars represent group means and vertical bars represent 95% confidence interval.

Same-Different Task

Next, we investigated whether CP effects are evident in the performance of same-different judgments and whether they are enhanced in the label condition. This entailed examining whether accuracy rates were greater for detecting between-category differences than within-category differences. To this end, we examined a subset of the data from the same-different task including only trials in which differing orientations were presented. We fitted same-different judgment accuracy to fully interacted fixed effect terms of block, labeling, and pair type (i.e., whether the displayed pair of aliens belonged to different categories or the same one). We also included a separate interaction term for block and number of trials per block (256 or 128) to account for variance in number of trials encountered.[3] Random intercepts were fitted for individual participants. An LRT was conducted to evaluate fixed effects(see Figure 5).

We found that overall accuracy improved throughout the task, as indicated by a main effect of block $\chi^2(2)=387.70$, $p<.001$. This effect is modulated by labels, such that, across

[3]The interaction between block and number of trials per block was revealed to be significant, $\chi^2(2)=67.90$, $p<.001$. Tukey pairwise comparisons indicate that, in block one only, those who encountered 256 trials achieved significantly higher odds of making accurate judgments than those who saw half as many, $p<.001$.

pair types, accuracy increased at different rates in label and no label conditions $\chi^2(2)=29.36$, $p<.001$. We also found that participants have higher odds of accurately identifying between-category differences ($M_{accuracy}=.79$, $SD=.13$) than within-category differences ($M_{accuracy}=.77$, $SD=.15$) overall, as revealed by a main effect of pair type, $\chi^2(1)=6.88$, $p=.009$. No main effect of labeling ($p=.981$) or number of trials per block ($p=.609$) was revealed.

CP effects were revealed to be larger in the label condition than the no label condition across blocks, as indicated by a significant interaction between labeling and pair type $\chi^2(1)=6.29$, $p=.012$. Tukey pairwise comparisons confirm that the direction of this effect indicates representational warping characteristic of CP in the label condition, such that odds of accurately detecting between-category differences ($M_{accuracy}=.80$, $SD=.16$) were significantly greater than odds of detecting within-category differences ($M_{accuracy}=.76$, $SD=.15$), $p<.001$. In the no label condition, the difference between the odds of accurately detecting between-category differences ($M_{accuracy}=.788$, $SD=.11$) and those of detecting within-category differences ($M_{accuracy}=.786$, $SD=.14$) were significantly smaller. No significant difference between pair types was revealed in the no label condition, $p=.933$. We find no significant two-way interaction between block and pair type ($p=.589$) or three-way interaction between block, labeling, and pair type ($p=.355$).

Post-hoc tests examining the interaction between labeling and pair type in individual blocks revealed that CP effects were significantly stronger in the label condition than the no label condition solely in block B $\chi^2(1)=5.54$, $p=.019$, not in block A ($p=.523$) or block C ($p=.178$).

Representational Similarity Analysis

The behavioral analyses showed larger CP effects in the label condition than the no label condition across blocks of the same-different task. However, post hoc tests examining this effect on a block-by-block basis revealed that the only single block within which this effect is significant is block B, where we find a significant CP effect in the label condition but none in the no label condition. For this reason, we focus our RSA analysis on blocks three and four of the approach-avoid task, which respectively precede and follow block B of the same-different task in which labels were associated with stronger CP effects. Using RSA, we evaluated whether perceptual representations of individual stimuli were modulated by labels in a pattern consistent with CP enhancement on the timescale of individual segments of time within the epoch.

To visualize these results, we subtracted between-category representational similarity scores from within-category in each of the labeling conditions and plotted these differences at each segment, resulting in a quantification of how much more similar representations of within-category pairs are to one another than representations of between-category pairs. Given that the objective perceptual similarity between these conditions is balanced, this can be considered a measure of

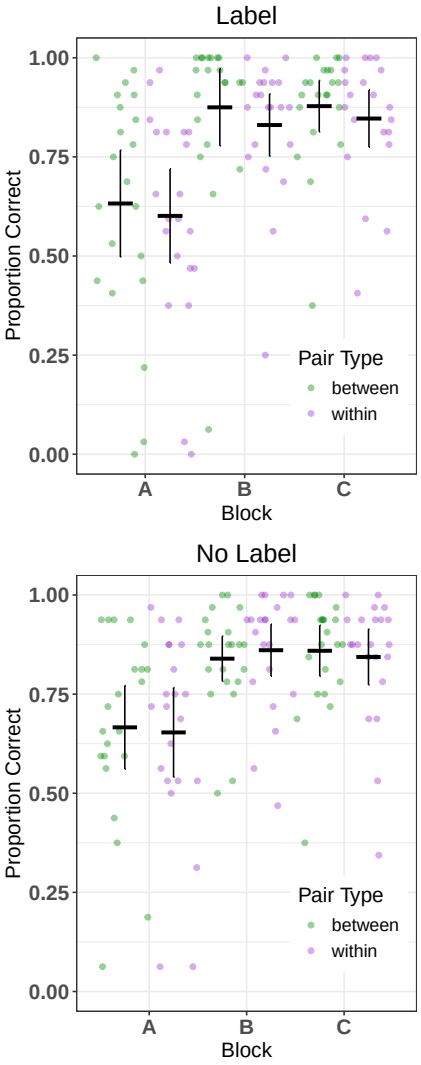

Figure 5: Plots illustrating larger CP effects, indicated by higher between-category accuracy than within-category accuracy, in the label condition (top) compared to the no label condition (bottom). Horizontal bars represent group means and vertical bars represent 95% confidence interval.

representational warping, which we will refer to as a CP index (CPI; see Figure 6). Since each of the curves plotted represents a difference between pair type conditions, the difference between these curves represents the interaction between labeling and pair type.

We calculated ηp^2 effect sizes of the interaction between labeling and pair type, representing the label-enhanced CP effect, for each segment of each block (denoted by the grayscale bars included in Figure 6), revealing that largest effect sizes of the interaction between labeling and pair type in block three occur in consecutive time segments beginning in the 540-580 ms post-stimulus onset window ($\eta p^2=.256$) and continuing in the 560-600 ms window ($\eta p^2=.249$) as well as the 580-620 ms window ($\eta p^2=.279$). ANOVA tests reveal that the in-

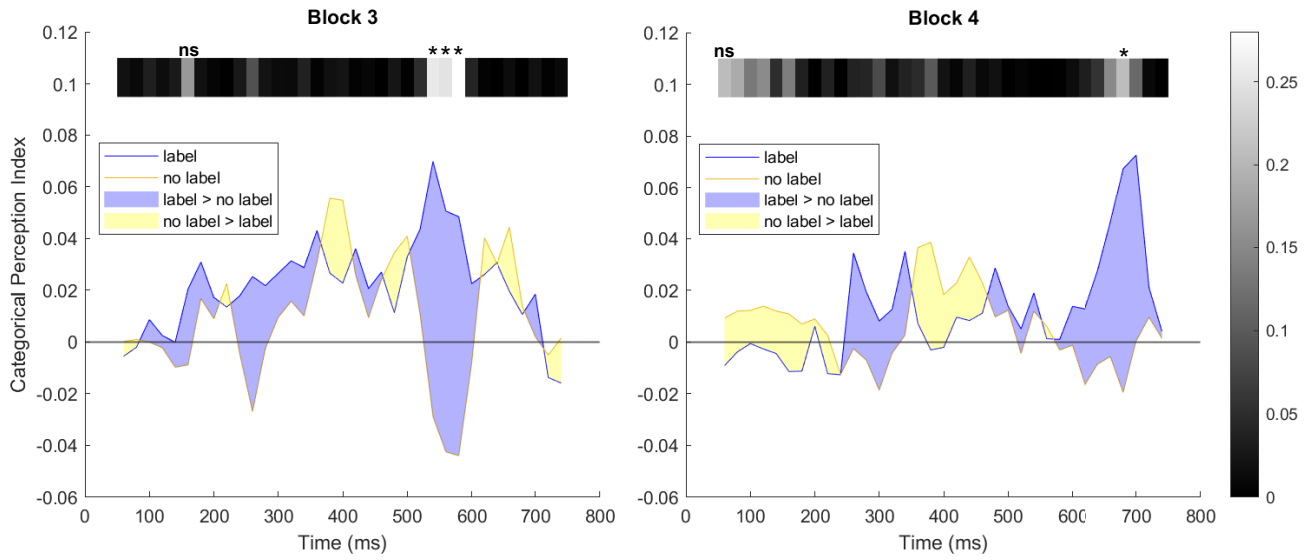

Figure 6: CPIs for the label and no label conditions in blocks three (left) and four (right) of the approach-avoid task. The grayscale bars at the top of each figure represent ηp^2 effect sizes of the interaction between labeling and pair type. Asterisks indicate significance at the 95% confidence level and 'ns' indicates nonsignificance where testing was conducted.

teraction was significant at each of these windows: 540-580 ms $F(1,76)=5.80$, $p=.019$, 560-600 ms $F(1,76)=5.67$, $p=.02$, 580-620 ms $F(1,76)=5.80$, $p=.019$. These findings suggest that, within this window, labels significantly enhance the label-enhanced CP effect such that CPI of the label condition is significantly greater than the CPI of the no label condition. The next largest effect size occurs 160-200 ms post-stimulus onset ($\eta p^2=.167$), but an ANOVA reveals the interaction is not significant during this window, $p=.102$.

In block four, the largest effect of the interaction between label and pair type occurs 680-720 ms post-stimulus onset ($\eta p^2=.167$), which an ANOVA reveals to be significant $F(1,76)=4.64$, $p=.034$. As in block three, this significant interaction indicates a significantly larger CPI in the label condition than the no label condition. The next largest effect size occurs in the 60-100 ms window ($\eta p^2=.206$), but does not reach significance, $p=.061$.

Discussion

The aims of the present exploratory study were to test whether labels facilitate category learning and to examine how labels influence the formation of representational warping patterns that are characteristic of CP. We hypothesized that category learning would be facilitated by labels and that CP effects, measured by representational warping, would be more prominent in the label condition. The behavioral results support this hypothesis, showing that labels facilitate category learning in the approach avoid task and that, overall, patterns of representational warping in the same-different task are more pronounced in the label condition than the no label condition. Further investigation on a block-by-block basis revealed that the only single block in which this effect was apparent was block B. Block B of the same-different task occurs at the halfway point of of the approach-avoid task, when the labeling advantage begins to become apparent. These results support our hypothesis that CP effects are strengthened by the presence of labels, particularly in the window during which the labeling advantage is most prevalent.

Suggestive evidence from the RSA provides converging evidence and additional insights regarding the relationship between the labeling advantage and CP. CPIs of stimulus representations indicate that labels are associated with significant enhancement of representational warping effects that are consistent with CP during select periods within the epoch. Specifically, we find significant modulation occurring around 500 ms post-stimulus onset and later. The fact that this modulation occurs at this stage of the epoch suggests that labels influence relatively late stages of perceptual analysis, possibly involving categorical evaluation or rule-updating, rather than lower-level perceptual processes. More work is required to determine whether additional category training with labels can result in modulation of other levels of perceptual analysis.

We were intrigued to observe that enhanced patterns of CP appear to slightly precede the labeling advantage. This may simply suggest a discrepancy in the sensitivity of our measures. However, it could also suggest a causal role of representational warping in the labeling advantage effect. Investigation on a more granular timescale is required to further evaluate this interesting possibility.

This study provides evidence that the increased rate of category learning precipitated by labels is accompanied by a corresponding enhancement of representational warping consistent with CP. These findings represent a step towards a richer understanding of the role of labels in category learning.

Acknowledgments

We thank undergraduate assistant Koa Rashidi for his assistance with data collection.

References

Delorme, A., & Makeig, S. (2004). Eeglab: An open source toolbox for analysis of single-trial eeg dynamics including independent component analysis. *Journal of Neuroscience Methods*, *134*, 9-21. doi: 10.1016/j.jneumeth.2003.10.009

Fotiadis, F. A., & Protopapas, A. (2022). Immediate and sustained effects of verbal labels for newly-learned categories. *Quarterly Journal of Experimental Psychology*. doi: 10.1177/17470218221126659

Goldstone, R. L., & Hendrickson, A. T. (2010). Categorical perception. *Wiley Interdisciplinary Reviews: Cognitive Science*, *1*, 69-78. doi: 10.1002/wcs.26

Heo, M., & Leon, A. C. (2005). Performance of a mixed effects logistic regression model for binary outcomes with unequal cluster size. *Journal of Biopharmaceutical Statistics*, *15*, 513-526. doi: 10.1081/BIP-200056554

Lopez-Calderon, J., & Luck, S. J. (2014). Erplab: An open-source toolbox for the analysis of event-related potentials. *Frontiers in Human Neuroscience*, *8*. doi: 10.3389/fnhum.2014.00213

Lupyan, G., Rakison, D. H., & Mcclelland, J. L. (2007). Language is not just for talking redundant labels facilitate learning of novel categories. *Psychological Science*, *18*, 1077-1083.

Mertens, A. J., & Colunga, E. (2024). Labels aid in the more difficult of two category learning tasks: Implications for the relative diagnosticity of perceptual dimensions in selective attention tasks. In *Proceedings of the annual meeting of the cognitive science society* (Vol. 46).

Peirce, J., Gray, J. R., Simpson, S., Macaskill, M., Höchenberger, R., Sogo, H., ... Lindeløv, J. K. (2019). Psychopy2: Experiments in behavior made easy. *Behavior research methods*, *51*, 195-203.