Labels aid in the more difficult of two category learning tasks: Implications for the relative diagnosticity of perceptual dimensions in selective attention tasks

Andrew J. Mertens (Andrew.Mertens@Colorado.Edu)
Department of Psychology & Neuroscience, 1905 Colorado Avenue, 345 UCB
Boulder, CO 80309 USA

Eliana Colunga (Colunga@Colorado.Edu)
Department of Psychology & Neuroscience, 1905 Colorado Avenue, 345 UCB
Boulder, CO 80309 USA

Abstract

Language represents a framework used to organize the things we experience. Redundant linguistic category labels facilitate category learning at a faster rate than category learning without labels (Luypan et al., 2007) suggesting language is also meaningfully involved in forming new categories. However, labels are not exclusively advantageous. Brojde et al., (2011) demonstrates that labels can be detrimental to category learning dependent on attending to historically agnostic dimensions over historically diagnostic ones (i.e., learning texture-based categories while ignoring shape). To separate historical experience from novel category learning, we task participants with classifying stimuli based on perceptual dimensions with less historical precedence as diagnostic cues for categorizing objects in everyday life (i.e., orientation and spatial frequency). Our results reveal a labeling advantage as well as slower overall learning in the orientation condition compared to spatial frequency-based learning. We discuss implications involving the historical use of these dimensions and the relationship between diagnostic and non-diagnostic dimensions.

Keywords: Labeling Effect; Category Learning; Selective Attention

Introduction

Language not only represents a method of conveying information and preserving ideas, but also a framework by which we organize the things we experience. The use of words to represent different groups of things inherently partitions them on the basis of key differences in some feature or collection of features. The distinction between a cup and a bowl, for example, will likely depend on a number of key features such as the diameter of the open side, the object's height, and the severity of the angle of the transition between the object's base and its sides. As in this example, shape characteristics are dominant as diagnostic features for determining membership in the majority of categories used to make sense of the physical world. Even in cases in which non-shape features are important for determining category membership, shape often still has a role to play. Cases in which shape has no bearing on classification are relatively rare, such as in the pure classification of substances (e.g., water, metal, cotton) or abstractions (e.g., blue, hunger, Tuesday). Reinforced by our language, it stands to reason that the featural norms we use to make categorization distinctions would influence the way we form new categories. In fact, as we will discuss below, there is evidence that the presence of linguistic labels during category learning has a direct, yet nuanced, role that may depend on the variation of both diagnostic and non-diagnostic sources of variance.

Background

Lupyan et al. (2007) demonstrate that categories given redundant labels are learned more quickly than unlabeled categories. In that experiment, participants were asked to distinguish between two groups of extraterrestrial species represented by images of complex three-dimensional models. One group of these aliens was arbitrarily characterized as friendly and approachable, the other being hostile and dangerous. Differences in certain sets of shape features of the models were indicative of category membership. Participants were tasked with learning how to classify the models based on trial-by-trial feedback. Critically, in this study category labels were redundant in that participants would only be presented with a category label representing the alien they had just seen after they had classified it and received feedback indicating whether their classification was correct or not. Despite the redundancy of labels in this design, participants in the labeling condition demonstrate faster category learning characterized by higher levels of accuracy compared to the 'no label' condition in early blocks of trials. These findings suggest that labels are involved in modulating attention to selectively accentuate diagnostic features of category membership.

The notion that linguistic labels modulate attention in ways that often benefit category learning has been reinforced in a number of studies, many of them in the developmental literature (for a review, see Sloutsky and Deng, 2019). For example, a study that measured the attention of infants using eye-tracking found that spoken words were associated with more rapid fixation on features that novel object exemplars had in common compared to conditions in which only the visual stimulus was presented or in which non-speech sounds accompanied the novel objects (Althaus & Mareschal, 2014). The spoken word condition was also the only one that proved to facilitate category learning.

Another line of evidence supporting the attention modulation explanation for the labeling effect comes from neuroimaging literature. In an electroencephalography (EEG) study, Maier and Abdel Rahman (2019) find a labeling effect in the N2 event-related potential (ERP) component after category training on novel objects. The N2 is a negative-going component that occurs approximately 150-200 ms after stimulus onset and is indicative of attentional target selection. The finding that category learning with labels is associated with a

In L. K. Samuelson, S. L. Frank, M. Toneva, A. Mackey, & E. Hazeltine (Eds.), *Proceedings of the 46th Annual Conference of the Cognitive Science Society.*

relatively larger N2 amplitude compared to training without labels is suggestive of the fact that training with labels causes a modulation of selective attention.

Notably, in all of the above demonstrations of the labeling advantage, shape is the sole diagnostic criterion by which novel categories are learned. The strength of this choice of stimuli is its external validity, given that shape is the most common single dimension along which objects are categorized. From a young age, children exhibit a bias for attending to shape when learning names for categories of novel solid objects (Smith et al., 1992) and grouping them by shape when asked to generalize names to them (Imai et al., 1994). This bias is only strengthened as children develop and approach adulthood (Landau et al., 1988). The shape bias has also been shown to generalize to Japanese despite the fact that the language's syntax does not make the count/mass distinction found in English (Imai & Gentner, 1997). Given this deeply engrained bias linking shape and category membership, it makes sense that labels would benefit the learning of shape-based categories.

The influence of labels is less straightforward when we examine category learning based on other, less historically dominant perceptual features. Brojde et al. (2011) examine the effect of labels in a category learning task in which the surface characteristics (i.e. color or texture) represent one dimension of variance in the stimuli and shape characteristics represent another. They found that labels were detrimental when learning categories on the basis of color or texture characteristics while ignoring shape: learning rates for this task were significantly higher in the condition that did not include labels. This deleterious labeling effect may have as much to do with the category-relevant dimension of this task as the irrelevant one. That is, when a feature as dominant in terms of diagnosticity as shape is irrelevant for categorization, it may be the case that labels cannot help but draw attention to that historically diagnostic but irrelevant feature, presumably in lieu of one that is relevant to category learning, which ultimately hinders learning.We explore this potential explanation by tasking participants with learning to categorize novel stimuli by selectively attending between dimensions that are historically uncommon for diagnosing category membership, namely orientation and spatial frequency, while holding shape variance constant.

Current Study

Motivated by the literature summarized above, this study seeks to examine the influence of labels on category learning when shape is held constant and only perceptual dimensions that have little to no history of being diagnostic of category membership are sources of variance in the stimuli. We use stimuli that vary independently on dimensions of spatial frequency and orientation in a category learning task that requires participants to learn to selectively attend to one dimension over the other.

A possible critique of past investigations of the labeling advantage is that the number of individual exemplars in each category is relatively small in some cases. The stimuli used in Lupyan et al. (2007) and Brojde et al., (2011) included the same 16 individual alien models. Over the course of hundreds of trials, it is likely that participants in these cases begin to recognize individuals as belonging to particular categories rather than depending solely on abstracted diagnostic features. Labeling effects resulting from learning such small categories are not necessarily any less meaningful because of this possibility, but for our purposes it is important to use large enough categories to be able to reasonably claim that participants are learning about general features that determine category membership.

To increase the size of the stimulus pool, we adapt the stimuli used by Tolins and Colunga (2015), a study that uses the Lupyan et al. (2007) paradigm to test how labels affect one's ability to adapt when categorization rules suddenly change. In it, participants learn to categorize a set of 36 stimuli with independently manipulable dimensions of spatial frequency and orientation. These stimuli not only allow us to expose participants to a relatively large number of unique stimuli, but also allow more control over variance in relevant and irrelevant dimensions.

Using our expanded stimulus set of Gabor patches with independently varying dimensions of spatial frequency and orientation, we investigate the influence of labels on category learning. We hypothesized that, in the absence of varying shape features for labels to point to, labels would facilitate higher rates of learning when category membership depended on either orientation or spatial frequency despite the fact that neither is historically diagnostic.

Method

Participants

Ninety-six undergraduate students between 18 and 23 years of age (M=19.36, SD=1.29) participated in the experiment in exchange for course credit. Forty-two participants reported identifying as female, 50 identified as male, three identified as non-binary, and one identified as gender non-conforming. Seventy-one participants identified as White, six identified as Hispanic or Latino, four as American Indian or Alaskan Native, four as Asian, two as Black or African American, two as Native Hawaiian or Pacific Islander, six as multiple or unlisted ethnicities, and one opted not to answer.

Stimuli

This experiment uses an expanded version of the set of stimuli used in Tolins and Colunga (2015). The stimuli consist of Gabor patches, gaussian-masked sinusoidal gratings, superimposed onto an unvarying image representing an alien body (see Figure 1). Thirty-six unique Gabor patches were used in the original Tolins and Colunga (2015) experiment, the lines of each varying between six possible spatial frequencies and six possible orientations. The Gabor patches in this experiment also vary in orientation and spatial frequency, with eight

values on each dimension, resulting in 64 possible combinations (see Figure 2).

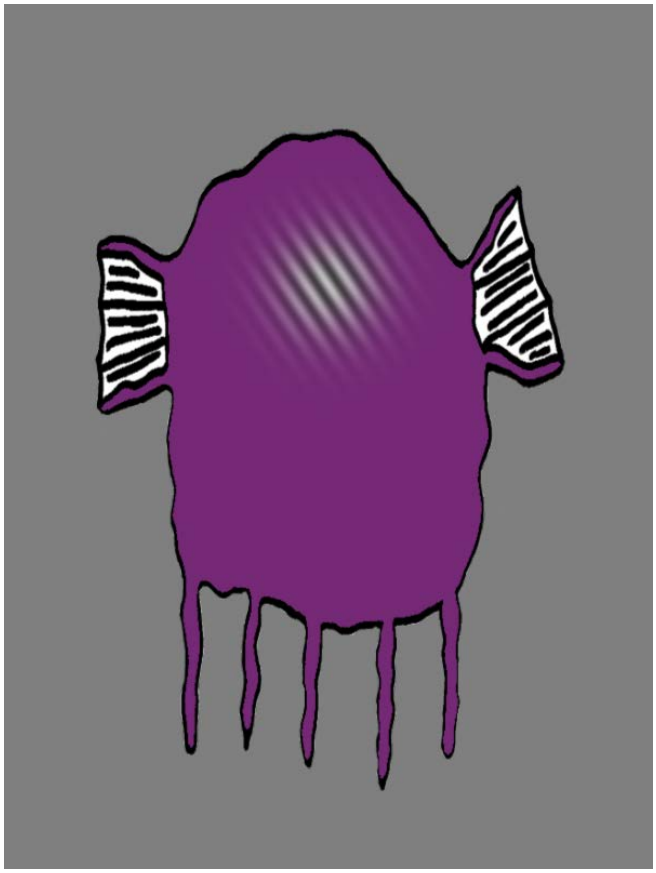

Figure 1: An example of the alien stimuli used in the category learning task.

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. These particular values were chosen based on the work of Campbell et al. (1970), who report that the just noticeable difference between spatial frequencies in the 2-6 cpd range are relatively stable at about 1.08 times the lower frequency. Therefore, by separating our spatial frequencies by a factor of 1.2 within the 2-6 cpd range we ensure that each step represents a roughly equivalent and readily perceptible difference in perceptual space. Categories based on spatial frequency divided these frequencies evenly, with the lowest four frequencies corresponding to one category of aliens and the higher frequencies corresponding to the other category.

The eight orientations used to create the stimuli described above consist of four orientations centered around the vertical orientation (150, 170, 10, and 30 degrees) and four centered around the horizontal orientation (60, 80, 100, and 120 degrees).

Design

Participants were trained on a novel object categorization task in which they classified each stimulus into one of two groups: hostile or friendly aliens. The relevant dimension for categorization, spatial frequency or orientation, was randomly assigned so that half of participants learned to classify aliens based on spatial frequency while ignoring variations in orientation and the other half learned orientation-based categories while ignoring spatial frequency variation. Participants were randomly assigned such that half only received feedback indicating whether their classifications were correct or incorrect (nonlabeled training) and the other half heard redundant pseudoword labels corresponding to the category they saw in addition to the feedback (labeled training). The pairings between pseudoword labels and categories were counterbalanced across participants in the labeled training condition. Participants were assigned to conditions that determined which subset of the stimuli they would see. In one condition, two thirds of the aliens participants saw represented the hostile category. Another condition presented the converse composition with two thirds of the aliens being friendly and, in a third condition, participants saw equal numbers of both trial types. All participants saw a subset of 36 of the total possible 64 combinations.

Procedure

Participants completed category training using stimuli adapted from Tolins and Colunga (2015). The task was designed using Psychopy behavioral testing software (Pierce et al., 2019). Data were collected using Psychopy version 2022.1.2 on Macintosh computers installed with Mohave version 10.14.6. The stimuli were viewed on built-in 1920- by 1080-pixel displays with brightness set to 50% from a viewing distance of approximately 60 cm. The full experiment took the average participant approximately 20-25 minutes to complete.

During category training, we first present participants with instructions explaining that they were to partake in a "training program to differentiate between extraterrestrial species on a newly found planet". They were also informed that they would be categorizing friendly and hostile aliens and that to differentiate them they would have to examine each alien's eye but were not told anything more about how category membership would be determined.

Each participant completed four blocks of trials, in each of which participants made category judgements of the same 36 unique stimuli. Within each trial of the initial training phase, an alien stimulus and an illustration of a space explorer are presented. The explorer can appear in one of four locations relative to the alien: above, below, to the left, or to the right. The participant is tasked with classifying each alien as friendly or hostile by either 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. For example, if the explorer appears to

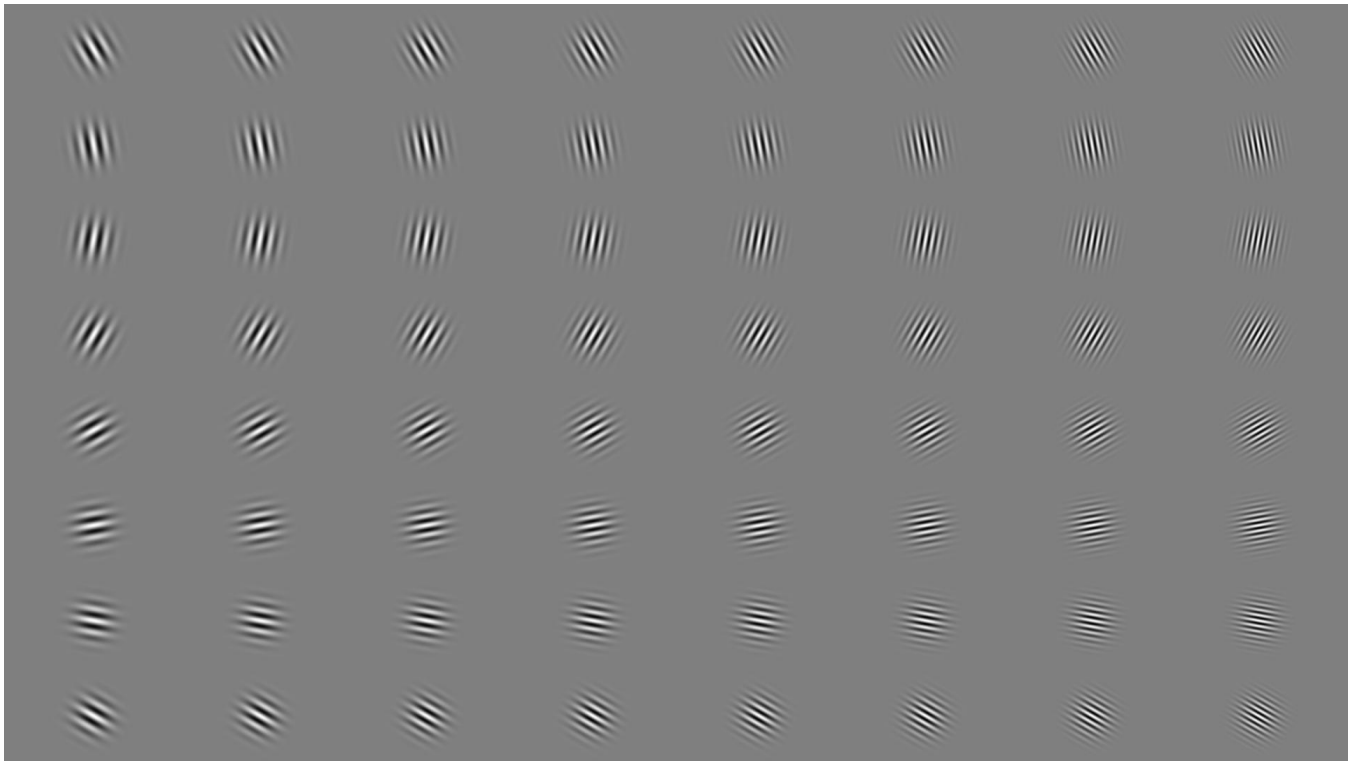

Figure 2: The 64 unique Gabor patches represented by the combinations of the eight selected spatial frequencies and orientations we use to generate our stimulus pool.

the right of the alien and the participant deems them friendly, they should press the left arrow key. After making a response, feedback is presented in the form of an auditory 'bloop' sound, in the case of a correct response, or a 'buzz' sound, in the case of an incorrect response. In addition, those in the labeled condition would hear the pseudoword label associated with the viewed category immediately following the feedback stimulus. These labels consisted of auditory recordings of a male voice saying gowachi /gOh-wOch-ee/ and havnori /hAv-nOR-ee/.

Results

Data from 35 participants is excluded from these analyses on the basis that their accuracy in the final block of training did not significantly exceed chance at an alpha level of .05. Exactly half of those assigned to the orientation condition, 24 of 48, did not achieve this benchmark. This data exclusion criterion is more restrictive than the one used in Tolins and Colunga (2015), in which only participants with mean accuracy scores under 50% were excluded. We did not use this criterion because we wanted to examine the influence of labels among those who demonstrated that they had learned how to reliably distinguish between categories. Even so, we ran the same analyses reported below using this alternative criterion, excluding 16 participants, and found the same pattern of effects.

To examine influences on category learning, trial accuracy is fit to a mixed effects logistic regression model. The main effects of training block, labeling condition, and relevant dimension are included as fixed effects in addition to the interaction effects between them. Factors representing participants, our 8 spatial frequency values, and our 8 orientation values were included as random effect factors. Initially, we tested this model with our three-level label condition factor, which included two conditions that included labels associated with categories in two different configurations (i.e., one condition in which the gowachi label was paired with hostile stimuli and havnori with friendly stimuli and the other with the opposite pairings). There was no significant main effect of label ($p = .402$) and examination of pairwise comparisons of means between the two conditions presenting labels within each block revealed no significant differences so we collapsed across them into a single 'label' condition. There was no effect of different compositions of trial types (i.e., mostly hostile, mostly friendly, and equal frequency conditions) so that factor was dropped from the analysis as well.

Our analysis reveals a main effect of block indicating that accuracy significantly improves over the course of training, $\chi^2(3) = 384.26$, $p < .001$. We also find a main effect of dimension indicating a significantly higher rate of accuracy when spatial frequency is diagnostic of category membership compared to orientation-based category learning, $\chi^2(1) = 6.91$, $p = .009$. There is no main effect of label condition, $p = .386$. A significant interaction between block and la-

bel $\chi^2(3) = 10.53$, $p = .015$ indicates that accuracy improves more quickly in the label condition than the no label condition. A significant interaction between block and dimension $\chi^2(3) = 15.52$, $p = .001$, indicates that accuracy improves more quickly in the spatial frequency condition than the orientation condition. We also find a significant three-way interaction between block, label, and dimension $\chi^2(3) = 9.14$, $p = .027$, indicating that the labeling advantage evident in the orientation condition is significantly greater than that of the spatial frequency condition. There is no interaction between factors of label and dimension, $p = .404$. Post-hoc analyses examining the label by block interaction separately by dimension indicate that there is a significant labeling advantage in the orientation-based categorization condition $\chi^2(3) = 14.11$, $p = .003$ but no significant labeling effect present in the spatial frequency-based categorization condition, $p = .211$. in The plot shown in Figure 3 illustrates the full pattern of results represented by these findings.

For a more granular examination of the timescale of label-augmented category learning we conducted additional logistic regression analyses on accuracy between each adjacent pair of blocks. Between blocks one and two, a pattern similar to the one described above emerges: we find a pronounced main effect of block, $\chi^2(1) = 103.40$, $p < .001$, a main effect of dimension, $\chi^2(1) = 6.40$, $p = .011$, an interaction between block and label, $\chi^2(1) = 5.79$, $p = .016$, an interaction between block and dimension, $\chi^2(1) = 7.69$, $p = .006$, and a three-way interaction between block, label, and dimension, $\chi^2(1) = 4.14$, $p=.042$. Between blocks two and three, significant main effects are revealed only for block $\chi^2(1)=14.64$, $p < .001$ and dimension $\chi^2(1) = 7.13$, $p = .008$. Lastly, between blocks three and four, a main effect of block $\chi^2(1) = 25.40$, $p < .001$ is revealed as well as an interactions between block and label, $\chi^2(1) = 4.43$, $p = .035$ and between block and dimension $\chi^2(1) = 7.55$, $p = .006$. This pattern of results from block to block indicates that the emergence of the labeling effect occurs early in training and persists until the final block when it finally dissipates.

Discussion

Previous work has shown that labels can be detrimental when category learning requires one to ignore an irrelevant but historically diagnostic categorical dimension like shape in favor of less typical diagnostic dimensions like texture and color, (Brojde et al., 2011). We raise the possibility that labels may be able to support category learning for uncommon diagnostic dimensions in cases where those dimensions are not pitted against a dominant, category-irrelevant dimension like shape. Our finding that labels aid in learning to categorize based on orientation while ignoring spatial frequency supports this notion. By holding constant the shape information inherent in our stimuli, we prevent the shape dimension from dominating attentive processes during category learning. Thus, at worst, labels have no effect on category learning in our experiment rather than hindering it.

Unexpectedly, labels facilitated an advantage for category learning in only one version of our category learning task: labels help when category membership depends on orientation, with spatial frequency being irrelevant. When the categories to be learned are distinguished by spatial frequency, with orientation being irrelevant, labels have no effect. The lack of a labeling advantage in spatial frequency-based category learning may be due to this version of the task being relatively easy resulting in a ceiling effect. Brojde et al., (2011) conclude that a ceiling effect is present when they report a similar finding that labels do not aid in a category learning task in which shape is the diagnostic feature. The simplest explanation for our finding is that the perceptual differences between stimuli were more difficult to perceive in the orientation condition of our experiment. Although we did not intend to vary the difficulty of these conditions, we also did not conduct extensive pilot testing to ensure the difficulty was balanced so this inequality is not entirely surprising.

This work reveals some key findings about which circumstances facilitate label-augmented category learning and which do not. First, our findings suggest that it is important to take both diagnostic and non-diagnostic perceptual dimensions into account when considering the effect of labels on category learning. Although labels may hinder learning when associated with a historically non-diagnostic dimension and pitted against a historically dominant diagnostic dimension like shape (Brojde et al., 2011), they can still be beneficial on a more level playing field in which both the diagnostic and non-diagnostic features are historically non-diagnostic. Our results also suggest that if the categorization task is too easy to begin with labels are not able to further benefit learning.

Future Work

Further work is necessary to fully understand the role that labels play in learning new categories. In an exit interview conducted in Brojde et al., (2011), participants who completed a version of the task in which categories were learnable by attending to shape, hue, or a conjunction of the two dimensions reported attending only to shape and ignoring hue when labels were present. This finding suggests that labels may play a role in drawing attention to individual dimensions at the expense of conjunctive dimensions. An additional line of research extending the present work could investigate this phenomenon of label-induced preference for single dimension selection in the context of the dimensions that are not historically used diagnostically, like the ones employed here.

An investigation of the modulation of attention during novel category learning might also improve our understanding of how labels are beneficial in this type of task. As mentioned previously, using EEG techniques Maier and Abdel Rahman (2019) find a labeling effect in an early ERP component indicative of attentional target selection after extensive category training on novel objects. To extend that work, examining ERPs during the course of a category learning task like the ones used here could provide meaningful insight into the time course and emergence of attentional mod-

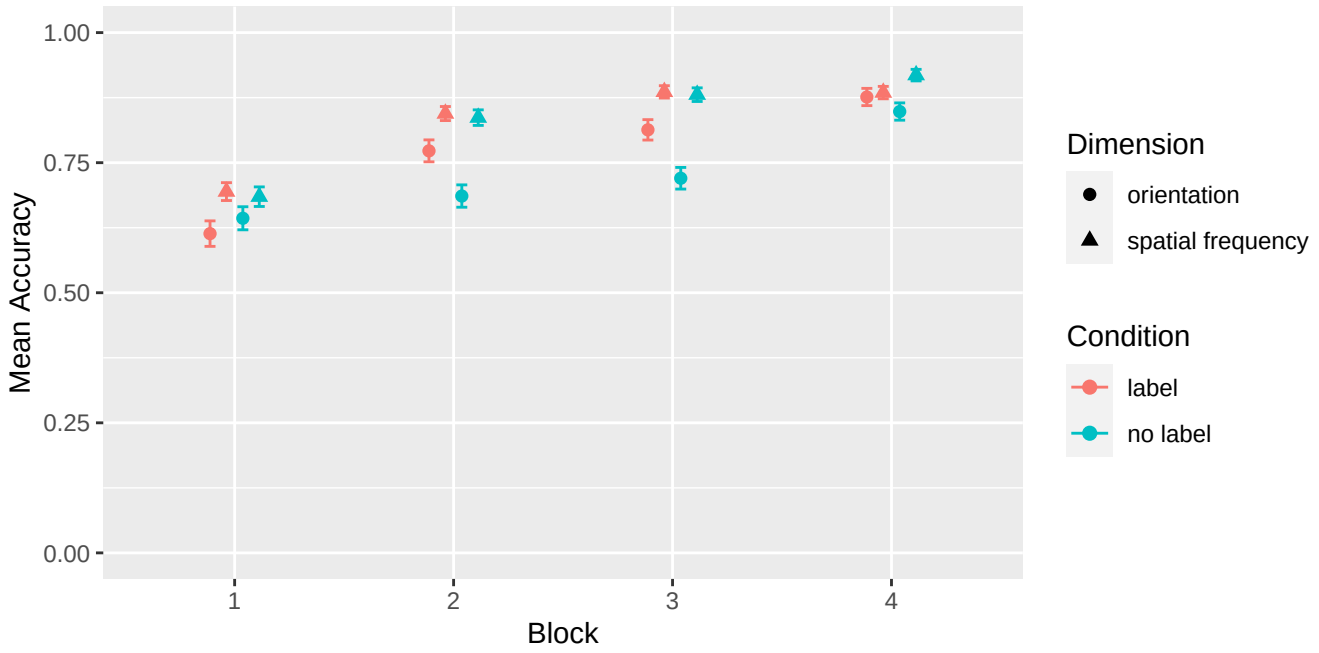

Figure 3: An illustration of the interactive effects of block, labeling condition, and relevant dimension on accuracy in category training. Error bars represent standard error.

ulation brought about by labels. Considering the brevity of the period during which the labeling advantage is evident, this could be achieved by interleaving a same-different task that requires participants to indicate whether pairs of stimuli are identical or not between blocks of category learning. These same-different trials would not only provide some indication of whether one version of the task is more difficult than the other at baseline but would also allow an examination of attentional modulation mid-learning when labels are apparently most influential. In combination with the work presented here, these lines of future research will begin to fill in some of the gaps in our understanding of how and under which circumstances words, as labels for categories, impact the way categories are learned.

Acknowledgments

We thank undergraduate assistants Shiva Ganesh Pandian and Koa Rashidi for their assistance with data collection.

References

Althaus, N., & Mareschal, D. (2014). Labels direct infants' attention to commonalities during novel category learning. *PLoS ONE*, *9*, 99670.

Brojde, C. L., Porter, C., & Colunga, E. (2011). Words can slow down category learning. *Psychonomic Bulletin and Review*, *18*, 798-804.

Imai, M., & Gentner, D. (1997). A cross-linguistic study of early word meaning: universal ontology and linguistic influence. *Cognition*, *62*, 169-200.

Imai, M., Gentner, D., & Uchida, N. (1994). Children's theories of word meaning: The role of shape similarity in early acquisition. *Cognitive development*, *9*(1), 45–75.

Landau, B., Jones, L. B. S. S. S., & Gleitman, L. (1988). The importance of shape in early lexical learning. *Cognitive Development*, *3*, 299-321.

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.

Maier, M., & Rahman, R. A. (2019). No matter how: Top-down effects of verbal and semantic category knowledge on early visual perception. *Cognitive, Affective, and Behavioral Neuroscience*, *19*, 859-876.

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.

Sloutsky, V. M., & Deng, W. (2019). Categories, concepts, and conceptual development. *Language, Cognition and Neuroscience*, *34*, 1284-1297.

Smith, L. B., Jones, S. S., & Landau, B. (1992). Count nouns, adjectives, and perceptual properties in children's novel word interpretations. *Developmental Psychology*, *28*, 273-286.

Tolins, J., & Colunga, E. (2015). How words anchor categorization: conceptual flexibility with labeled and unlabeled categories. *Language and Cognition*, *7*, 219-238.