*Language and Cognition* **7** (2015), 219–238. doi:10.1017/langcog.2014.26


# How words anchor categorization: conceptual flexibility with labeled and unlabeled categories*‡

JACKSON TOLINS

*Department of Psychology, University of California, Santa Cruz*

AND

ELIANA COLUNGA

*Department of Psychology and Neuroscience, University of Colorado, Boulder*

(*Received* 01 *December* 2013 – *Revised* 25 *May* 2014 – *Accepted* 06 *June* 2014 – *First published online* 22 *July* 2014)

ABSTRACT

Labeled categories are learned faster, and are subsequently more robust than categories learned without labels. The label feedback hypothesis (Lupyan, 2012) accounts for these effects by introducing a word-driven top-down modulation of perceptual processes involved in categorization. By testing categorization flexibility with and without labels, we demonstrate the ways in which labels do and do not modulate category representations. In Experiment 1, transfer involved a change in selective attention, and results indicated that labels did not impact relearning. In Experiment 2, when transfer involved a change in the behavioral response to categories whose structures did not change, a reversal shift, learning the categories with labels speeded recovery. We take this finding as evidence that the augmentation of perceptual processes by words is on the one hand fairly weak without explicit reinforcement, but on the other allows for category representations to be more abstract, allowing greater flexibility in behavior.

KEYWORDS: categorization, language and cognition, flexible cognition, selective attention

[*] Portions of this study were presented at the 35th annual meeting of the Cognitive Science Society, in Berlin, Germany. The authors thank the conference reviewers and attendees for their useful feedback. The authors also thank the many research assistants at the CU Language Project who assisted in this project. Address for correspondence: Jackson Tolins, University of California, Psychology Department, 1156 High St., Santa Cruz, CA 95060. e-mail: jtolins@ucsc.edu

[‡] The original version of this article was published with last word of the title missing. A notice detailing this has been published and the error rectified in the online and print PDF and HTML copies.

## 1. Introduction

Does having a word for something change the way we think about it? Language, along with its use in communication and social interaction, provides a symbolic system of representation through which a speaker contemplates the world around them. The emergence of the capacity for symbolic representation transformed human cognition (Deacon, 1997; DeLoache, 2004), permitting abstract thought and making possible cultural transmission of knowledge. Yet the relationship between language and other cognitive processes is still controversial. For many who view language as a distinct mental module (Gleitman & Papafragou, 2005; Pinker, 1995), language is merely a formal medium that is used to describe mental representations, while remaining independent of the concepts it expresses (Li & Gleitman, 2002). Recent work in understanding what happens when categories are learned with words, and when individual objects are labeled, challenges this division. Instead, the relationship between the perceptual processes in categorization and the concepts' representations appears to be mutually supportive and bi-directional (Bowerman & Choi, 2001; Goldstone, Lippa, & Shiffin, 2001; Lupyan, 2012). The role of words in shaping conceptual development has been well explored in the literature (Casasola, 2005; Gentner & Goldin-Meadow, 2003; Gumperz & Levinson, 1996; Levinson, 1997; Lupyan, Rakison, & McClelland, 2007; Spelke & Tsivkin, 2001; Waxman & Markow, 1995; Yoshida & Smith, 2005). A lively discussion currently exists in the field as to the extent of this causal relationship, both in early development and adulthood.

The studies presented here are focused on the role of verbal labels in shaping perception of object categories. Research in this domain has demonstrated that words influence category learning and representation, with those categories learned with words learned quicker and more robustly (Lupyan et al., 2007; Lupyan & Thompson-Schill, 2012). The LABEL FEEDBACK HYPOTHESIS (Lupyan, 2012) suggests that the effect of labels on categorization is due to the on-line, top-down influences of words on lower processes, including perception. Lupyan (2012) suggests that this modulation of perception is deeply penetrating, but functions on-line in a temporary manner rather than influencing learned categorical perception. This conceptualization of a label's feedback as deep yet transient is supported by a number of studies that demonstrate that verbal interference, removing linguistic influence, washes out any effect of labeling on categorical perception (Gilbert, Regier, Kay, & Ivry, 2006; Lupyan, 2008; Winawer, Witthoft, Frank, Wu, Wade, & Boroditsky, 2007).

Another means by which we can explore the role of verbal labels in categorization is through testing labels' role in conceptual flexibility. In the studies presented below, we combined the category-learning paradigm

(e.g., Lupyan et al., 2007) with a subsequent transfer, or shift task, in which participants had to learn to respond to objects in the same stimulus space in a novel manner. This transfer learning was done without any secondary tasks that would lead to verbal interference, leaving the language system, and any influence on categorization it may have, intact. Transfer learning tasks are useful in demonstrating the structural relationship between category representations: relearning tasks with highly overlapping mental representations should be easier to learn than those in which the categories are highly divergent (Hendrickson, Kachergis, Fausey, & Goldstone, 2012). Where the transfer profiles of categories learned with and without labels differ, differences would provide evidence of how verbal labels modulate category representations, giving us a better understanding of the ways that language influences concepts and categories during encoding and retrieval.

## 2. Categorical perception and flexibility

The categories we possess influence both our judgments of similarity and ability to discriminate distinct objects. Objects within a category are judged as being more similar than objects that do not share a category label (Goldstone, 1994; Goldstone et al., 2001). This categorical perception allows perceptually distinguishable stimuli to be treated as the same and responded to in kind (Harnad, 2005).

Importantly, these differences in judgment are partially driven by changes in low-level representational change and perception. A number of mechanisms underlie categorical perception, including attentional weighting, stimulus imprinting, differentiation, and unitization (Goldstone, 1998). Attentional shifts during categorization learning lead to psychological 'stretching' along those dimensions that are historically diagnostic for category membership (Nosofsky, 1986). This shift in weighting is tied to a de-emphasis on non-salient features, leading to acquired equivalence along these dimensions (Haider & Frensch, 1996; Honey & Hall, 1989).

The gradual perceptual warping that occurs during learning is long-term, as demonstrated by an effect of category learning on conceptual flexibility. Goldstone and Steyvers (2001) tested the role of attentional weighting on subsequent relearning, manipulating whether the previously relevant, previously irrelevant, or some novel dimension became relevant for categorization after an initial learning period. Changes in categorization of the same stimulus space that make use of the same relevant dimension were the easiest to relearn. Learning to pay attention to previously irrelevant dimensions was more difficult than learning categories based on dimensions that were not part of the original learning (Goldstone & Steyvers, 2001). These findings demonstrate that learned selectivity along a particular dimension continues to

capture attention when this dimension becomes irrelevant to the task at hand (see also Shiffrin & Schneider, 1977). Similarly, perceptual warping that reduces the perceptual space along irrelevant dimensions continues after learning, interfering with subsequent learning when these dimensions become relevant.

The examples of transfer presented above can be categorized as extra-dimensional shifts, in which the diagnostic dimension within the stimulus space changes, focus on the relationship between individual stimuli and category representation. This research, in combination with studies making use of reversal shift learning tasks, have lead researchers to posit a second level of association, that between category representation and category response (Kendler & Kendler, 1962; Kruschke, 1996; Maddox, Glass, O'Brien, Filoteo, & Ashby, 2010). During a reversal shift, the stimulus dimension that was relevant during category learning remains relevant, but the overt responses to the category reverse. Studies have demonstrated that reversal shifts are easier to learn than extra-dimensional shifts (Goldstone & Steyvers, 2001; Kruschke, 1996), suggesting that this type of shift allows for perception-to-category associations to remain intact while the category-to-response, or label (Maddox et al., 2010), association is changed.

Taken together, these studies demonstrate the utility of shift learning tasks in uncovering the underlying processes involved in categorical perception. This paradigm has yet to be applied to the study of verbal labels and cognition, and may prove to be useful for illuminating the degree to which labels modify or augment categorical processing during regular processing of objects, without taking the language faculty off-line through the additional implementation of verbal interference.

## 3. Verbal labels and categorization

For many researchers, verbal labels are simply that; names that get attached to categories while remaining separate from what they are used to represent (see, e.g., Hespos & Spelke, 2004). From this perspective, verbal labels may assist category learning by providing an opportunity for training and practice, but do not take part in modulating the perceptual processes involved in categorization themselves. In contrast, recent research has demonstrated that words directly influence these lower-level processes, modulating categorical perception.

The representation of objects in labeled categories appears to be distinct from those in non-labeled categories. For example, participants judged objects from labeled categories as being more similar to each other, and more distinct from contrasting categories (Goldstone, 1998; Goldstone et al., 2001). Is this similarity derived from the simple fact that the objects share a label, giving them one more thing in common, or does the presence of a label change the perception of the object itself? It appears that simply

sharing a category label is not sufficient to modulate within-category similarity, rather the presence of words during categorization and category learning highlight meaningful dimensions of the stimulus space (Boroditsky, Schmidt, & Phillips, 2003).

This theory is supported by a number of studies that have contrasted languages for which a particular stimulus domain is either divided into two labeled categories with languages that do not make the same distinctions in their own lexicon. The domain of color has been a particularly active, if controversial, field for the exploration of linguistic effects on perception, as languages vary in the number of color terms used to divide the visual experience. These differences in color categorization can cause effects on both memory for color and color perception (Gilbert et al., 2006; Roberson, Davidoff, Davies, & Shapiro, 2005; Winawer et al., 2007). For example, speakers of Russian, who make a lexical distinction between light and dark blue that English speakers do not, when asked to determine which of two colors were the same as a third responded more quickly when one of the colors was from a different lexical category than when both were of the same. This categorical perception effect was not present for English speakers, for whom the colors all belonged to the same category.

In this way, labels interact with category learning, making categories activated through words different than categories activated without: "by virtue of the learned associations between words and their referents, words participate in the creation of categories they denote, and function on-line to selectively shape the perceptual representations that underlie our conceptual knowledge" (Lupyan, 2007, p. 2). Activation of a category's label during perception results in a temporary heightening of attentional weighting, such that the perceptual space becomes more categorical in nature. In a task where participants were to identify the number 5 in a field of 5s and perceptually similar 2s, hearing the word 'five' prior to the presentation of the field improved performance (Lupyan & Spivey, 2010). Feedback from the verbal label augments perception, strengthening the visibility of the dimensions relevant for categorization while abstracting over irrelevant dimensions (Lupyan, 2008, 2009; Lupyan & Spivey, 2010).

The on-line nature of words' function in this manner appears to be key. The cross-linguistic differences in categorical perception can be extinguished with the addition of verbal interference; by removing the activation of category labels, those speakers whose languages make distinctions treat a given perceptual domain as would speakers from languages that do not make a lexical distinction. For example, English speakers who make a distinction between 'green' and 'blue', treat colors on either side as less similar than speakers of languages that do not. For both these speakers, and the Russian speakers described above, the categorical difference in processing these colors is eliminated when a secondary task removes the possibility of label activation (Kay & Kempton, 1984; Roberson & Davidoff, 2000). Thus, verbal labels do

appear to modulate lower-level perception, but in a transitory or impermanent manner, in contrast to the long-term influence of category learning itself.

We present a pair of experiments that test for similar effects through the use of shift learning tasks, which leave verbal processing free from interference while requiring participants to re-structure the stimulus space. Theories of shift learning have emphasized both selective attention as well as mediating responses or intermediate representations between stimulus perception and response selection (Krushke, 1996). Given that categorization seems to involve two levels of associations, with label selection being the proposed top tier, it is possible that the top-down influence on labels has differential influence on these two separate levels. By contrasting an extra-dimensional shift and a reversal shift learning task with and without category labels, we are able to distinguish between the influence of labels on category representation at the distinct levels of processing. The extra-dimensional shift task tests the influence of verbal labels on learned selective attention. Perception of stimulus from categories learned with labels should demonstrate heightened discriminability, while transfers in categorization that cross this learned perceptual emphasis, such as in changing the diagnostic dimension, should be more difficult when learned with labels. The reversal shift learning task tests the degree to which the presence of category labels modulates the association between category membership and overt response.

## 4. Experiment 1

In the first experiment, the transfer after training involved crossing learned selective attention, which we call the extra-dimensional shift task. This task sought to provide a measure of the degree to which on-line top-down influence of verbal labels during transfer impinges on relearning a novel categorization. In this transfer condition, participants learned to change from one diagnostic dimension of the stimulus space to another, cross-cutting category boundaries. This type of transfer crosses learned selective attention, which has a steep cost in relearning (Goldstone & Steyvers, 2001).

When an individual needs to restructure the categorical divisions of a particular domain, especially when this restructuring requires a shift in attention to a previously non-diagnostic dimension, having verbal labels for categories already established could slow down relearning. This could be for two different reasons. If labels act on categories by strengthening selective attention on a long-term scale, the dimensional switch transfer of labeled categories should have an increased cost compared to unlabeled categories. Similarly, the automatic activation of previously learned labels, along with their subsequent top-down influence on perceptual processes, during the presentation of objects during the transfer phase could inhibit relearning by perseveratively strengthening activation of the now irrelevant dimension.

## 4.1. METHOD

### 4.1.1. *Participants*

Sixty-five participants were drawn from the undergraduate psychology subject pool at the University of Colorado, Boulder, and participated in exchange for course credit. Participants were randomly assigned to either a labeled or unlabeled category learning condition.

### 4.1.2. *Materials*

Categories of 'aliens' were created as the stimulus set for this experiment. In order to create stimuli that varied along two dimensions, we made use of gabor patches, which vary both in the orientation of the lines in the patch, and the spatial frequency of these lines. Six points along each dimension, orientation, and frequency were chosen, creating 36 total gabor patches (see Figure 1). These patches were embedded in the stimuli as the aliens' eyes. The two categories that participants learned were organized based on the kind of eyes the aliens had.

### 4.1.3. *Training procedure*

Following the procedure from Lupyan et al. (2007), participants were told that they were to take part in a NASA training program before traveling to a newly found planet. In training, it was explained that previous explorers to the planet had discovered two similar looking aquatic alien species, one of which was friendly and could be approached, and one that was dangerous and had to be avoided. In the label condition, the participants were told that the explorers had decided to name the aliens, and that the friendly aliens were named 'Gowachi', while the dangerous aliens were named 'Caleba'. The training NASA provided was to teach the participants how to distinguish between the two. Thus, participants were asked to learn to distinguish between two categories within a set of novel stimuli. The categorization learned in the training phase separated the aliens by the spatial frequency of their eyes, with thick-banded aliens, the 'Gowachi', being friendly and approachable, and thin-banded aliens, the 'Caleba', unfriendly.

Individual trials began with a fixation marker in the middle of the screen, presented for 500 milliseconds. For each trial, an alien was presented alone in the center of the screen (500 ms), before a scuba diver joined the alien, appearing in one of four locations; above, below, or on either side of the alien (see Figure 2). The participant then decided whether to approach or escape the alien using the directional keys on a standard keyboard. For example, if a scuba diver appeared on the left of a friendly alien, the participant should press the 'right' key to move the scuba diver closer. If a scuba diver appeared above an unfriendly alien, the participant should press the 'up' arrow key to escape the alien. Participants had 3000 ms to respond once the diver appeared. After a response was made,

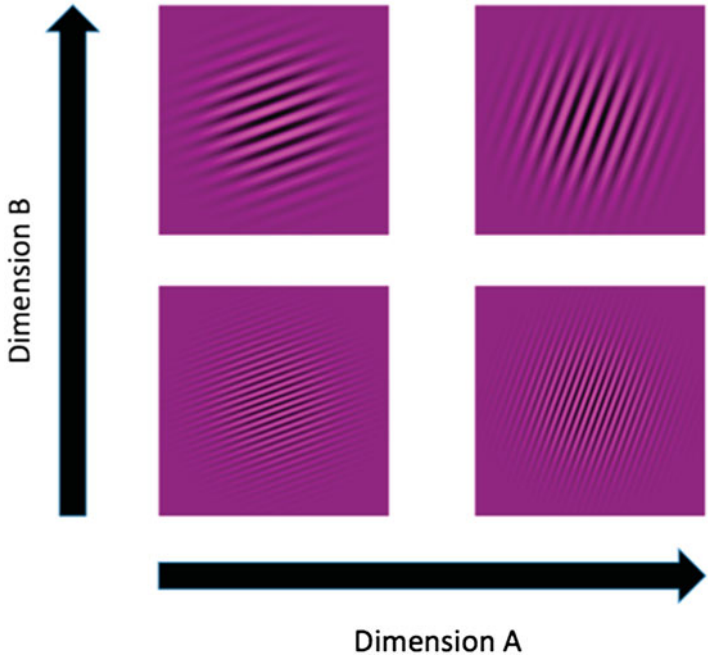


Fig. 1. The stimulus space varied along two dimensions: orientation (Dimension A) and spatial frequency (Dimension B), the boundaries of which are delineated by the four exemplars. 6 locations along each dimension were chosen, creating a total of 36 stimuli.

or the participant failed to provide a response in time, feedback was provided. In the no-label conditions, feedback was minimal (a chime for correct, a buzz for incorrect), while in the label conditions feedback included the correct label for the stimulus item (chime/buzz + spoken correct category label). Following feedback, the alien and scuba diver remained on the screen for additional 800 ms before the start of the next trial and the re-presentation of the fixation marker. There were four blocks of training. In each block, each of the 36 alien exemplars was presented once, in random order. Thus, over the whole training phase, each unique alien + diver trial was presented once, for a total of 144 trials of training (36 alien exemplars × 4 diver locations). All subjects received the same number of categorization learning trials and had equal exposure to the stimuli across conditions.

4.1.4. *Transfer procedure*

After training was complete the participants were told that they were now ready to travel to the Planet Teeb. Upon arrival on the planet the participants were alerted that something has gone wrong, and that the aliens are not

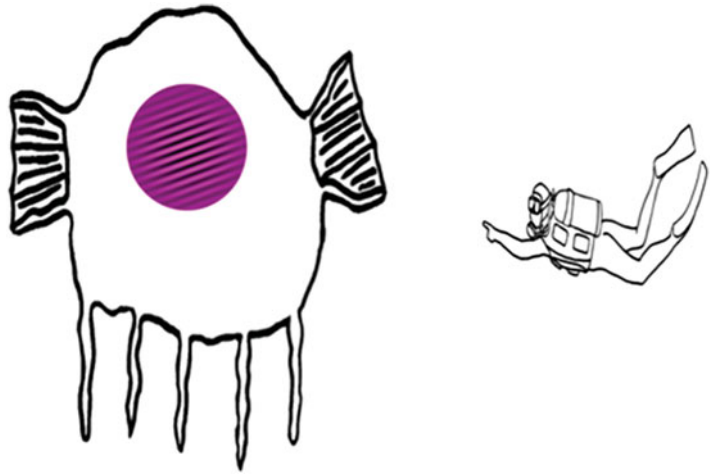

Fig. 2. Stimuli were presented as aliens whose eye patterns indicated category membership. For each trial, a diver would appear in one of the four cardinal directions, as shown in the figure. Participants were to decide whether the diver should approach or escape.

behaving as NASA had thought. Participants then faced a relearning task, in which the diagnostic dimension was changed, requiring a relearning of selective attention. Participants were not explicitly told about the nature of the change, just that categorization was different than expected. Participants who learned during training to pay attention to the spatial frequency of the lines in the eyes here had to learn to categorize the friendly and unfriendly aliens based on the steepness of the orientation of the bands, ignoring spatial frequency. This meant that half of each category learned during the first phase subsequently became part of the new category structure learned during transfer testing, or that half of the friendly aliens that were approached must now be considered unfriendly and so avoided, and the reverse. Similarly, in the label condition, half of the 'Gowachi' must now be treated as 'Caleba' and approached, and half the 'Caleba' as 'Gowachi' and avoided.

The post-transfer phase consisted of a second set of 144 trials, again presented in random order within each of the four presentation blocks. During the transfer phase trials only minimal feedback (chime or buzz) were given in all conditions, whether label or no-label. The full experiment took participants on average 25 minutes to complete.

### 4.2. RESULTS

Each correct trial was scored as 1, each incorrect trial as 0, and each trial in which the participant did not answer was dropped from the analysis. Accuracy across

each of the four training blocks and four transfer blocks was then calculated. The data from those participants who did not reach at least 50% accuracy by the end of training were not included (6 participants in total), leaving 31 in the label condition and 28 in the no-label condition. Data were then entered into a mixed factor repeated measures analysis of variance (ANOVA) with label as a between-subjects factor and block as a within-subjects factor, for each phase.

#### 4.2.1. *Training phase*

Analysis revealed a significant main effect of block ($F(3,171) = 50.92$, $p < .001$, partial $\eta^2 = .47$). Participants learned to correctly categorize the aliens over the course of the training phase, from an average 67% accuracy in the first block ($SD = 21\%$) to an average of 91% accuracy in the fourth and final block ($SD = 16\%$) (see Figure 3). There was no main effect of label ($F(1,57) = 0.15$, $p = .70$). Similarly, there was no significant interaction between block and label condition ($F(3,171) = 0.614$, $p = .61$).

#### 4.2.2. *Transfer phase*

Similarly to training, while participants' accuracy improved over the four blocks of transfer, from 53% ($SD = 12\%$) in the first block of transfer to 62% ($SD = 21\%$) in the final block, with a significant main effect of block ($F(3,171) = 8.48$, $p < .001$, partial $\eta^2 = .13$), there was no significant interaction between block and label condition ($F(3,171) = 0.68$, $p = .41$) (see Figure 3).

#### 4.2.3. *Transfer cost*

We tested whether the presence of a label during learning would affect the cost of transfer. This was conducted by comparing the change in accuracy from the last block of training to the first block of transfer across the label conditions. The cost of transfer in the labeled learning condition ($M = 39\%$, $SD = 15\%$) was not significantly different than the cost of transfer for the unlabeled learning condition ($M = 37\%$, $SD = 22\%$) ($t(57) = 1.3\%$, $p = .78$). The reduction in accuracy caused by the implementation of the transfer task was similar across conditions.

### 4.3. DISCUSSION

The presence of verbal category labels during learning did not reduce flexibility. While participants did learn to categorize based on the previously irrelevant dimension, this did not differ whether the original categories had

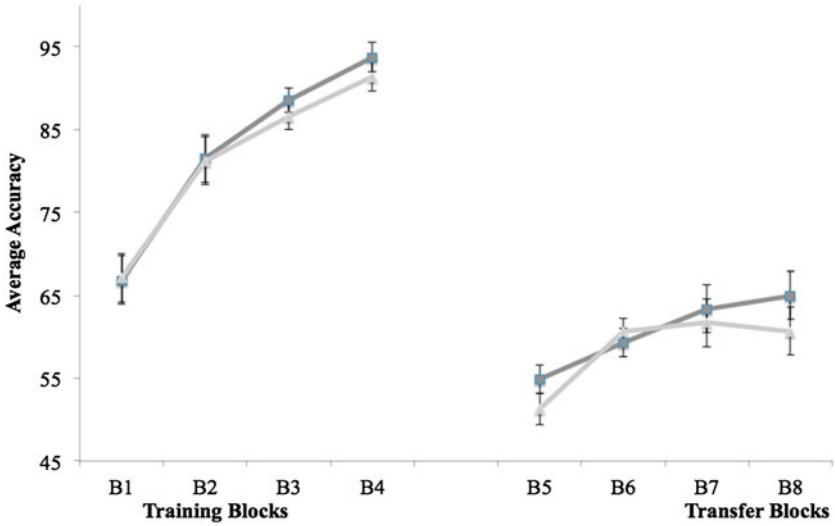


Fig. 3. Average accuracy by block in the labeled and unlabeled conditions for the training and transfer phases of the dimensional shift task. Error bars represent Standard Errors.

been learned with a label or not. This result does not support a perspective on the role of verbal labels as speeding or increasing selective attention over a long-term scale, which would have lead to decreased flexibility (Goldstone & Steyvers, 2001), but does provide convergent evidence along with verbal interference studies that the effects of labels on categorical perception is on-line and transient.

The label feedback hypothesis (Lupyan, 2012) suggests that after the explicit feedback during learning, labels are automatically activated upon the perception of an object, and this activation leads to top-down modulation of perceptual representations. If so, a strong version of this hypothesis would suggest that the activation of labels at transfer should have continued to draw attention to the previously diagnostic stimulus dimension, that of spatial frequency, reducing conceptual flexibility. This view is not supported by the current data. Given only minimal feedback at transfer, participants who originally learned to distinguish between 'Gowachi' and 'Caleba' had a similar cost of switching and relearned the categorization of the aliens at a similar rate to those that did not learn category labels.

## 5. Experiment 2

The previous experiment tested whether labels would influence conceptual flexibility when the categories themselves changed, requiring adjustment of

learned selective attention. Experiment 2 tested whether categories learned with or without labels influence conceptual flexibility at a different level, that of the association between category and response. To do this we used a reversal shift transfer task, in which participants had to switch their responses to previously learned categories, without changing the boundaries of the categories themselves.

This task tests the degree to which labeled categories are more or less abstract and flexible than unlabeled categories. Prior research suggests that categories activated by labels are more categorical in nature than those activated through other means (Lupyan & Thompson-Schill, 2012). From this perspective, the predictive top-down influence of labels acts as a stand-in for instances of category members, which would make the representations of the objects perceived more abstract (Clark, 2006; 2013; Lupyan et al., 2007). Support for this perspective would be found in a distinct pattern of recovery from the reversal shift transfer for labeled categories compared to categories learned without a label. Previous research on the actives–latent account of cognitive flexibility indicates that more abstract representations support behavioral switching (Cohen & Servan-Schreiber, 1992; Munakata, 1998; Kharitonova, Chien, Colunga, & Munakata, 2009). This suggests that labeled categories should provide the means for more flexible behavioral responses than unlabeled categories.

### 5.1. METHOD

#### 5.1.1. *Participants*

Fifty-three participants were drawn from the undergraduate psychology subject pool at the University of Colorado, Boulder, and participated in exchange for course credit. Participants were randomly assigned to either a labeled or unlabeled category learning condition.

#### 5.1.2. *Training procedure*

The training procedure was the exact same as in Experiment 1, except the categories learned divided the stimulus space along the dimension of orientation. During the training phase, participants learned to categorize based on the orientation of the lines that made up the aliens' eyes, with shallow-sloped lines, or 'low-orientation aliens', being friendly, and steep-sloped lines, 'high-orientation', unfriendly, while ignoring the dimension of spatial frequency.

As before, participants in the labeled condition received not only feedback on their accuracy for each trial, but also the correct category label for each alien.

#### 5.1.3. *Transfer procedure*

For the reversal shift transfer, the diagnostic dimension, line orientation, remained the same, as did the boundary location along this dimension, but the escape/approach responses were switched. Here, participants had to relearn that aliens that were previously approachable were now dangerous, and aliens that had been dangerous during training were now friendly and should be approached.

### 5.2. RESULTS

Again, each correct trial was scored as 1, each incorrect trial as 0, and each trial in which the participant did not answer was dropped from the analysis. Accuracy across block was then calculated. The data from those participants who did not reach at least 50% accuracy by the end of training were not included (7 participants in total), leaving 25 participants in the label condition and 21 participants in the no-label condition. Data were then entered into a mixed factor repeated measures analysis of variance (ANOVA) with label as a between-subjects factor and block as a within-subjects factor, for each phase.

#### 5.2.1. *Training phase*

Accuracy increased from 56% in the first block of training to 75% in the final block. Analysis revealed a significant main effect of block ($F(3,132) = 25.76$, $p < .001$, partial $\eta^2 = .37$). There was no main effect of label type ($F(1,44) = 0.002$, $p = .97$), nor was there an interaction between block and label ($F(3,132) = 1.15$, $p = .33$) (see Figure 4).

#### 5.2.2. *Transfer phase*

As in Experiment 1, there was a significant main effect of block during the transfer phase ($F(3,132) = 3.24$, $p < .05$, partial $\eta^2 = .07$). However, there was also a significant interaction between block and label type ($F(3,132) = 5.96$, $p < .01$, partial $\eta^2 = .12$). Post-hoc analyses revealed a significant interaction between block and label for each adjacent pair of blocks ($F(1,44) = 11.59$, $p = .001$; $F(1,44) = 18.1$, $p = .0001$; and $F(1,44) = 4.57$, $p = .038$, respectively).

#### 5.2.3. *Transfer cost*

Participants in the labeled learning condition had a similar cost of transfer from the last block of training to the first block of transfer ($M = 3.6\%$, $SD = 12\%$) to those in the unlabeled learning condition ($M = 1.3\%$, $SD = 12\%$) ($t(44) = .63$, $p = .53$).

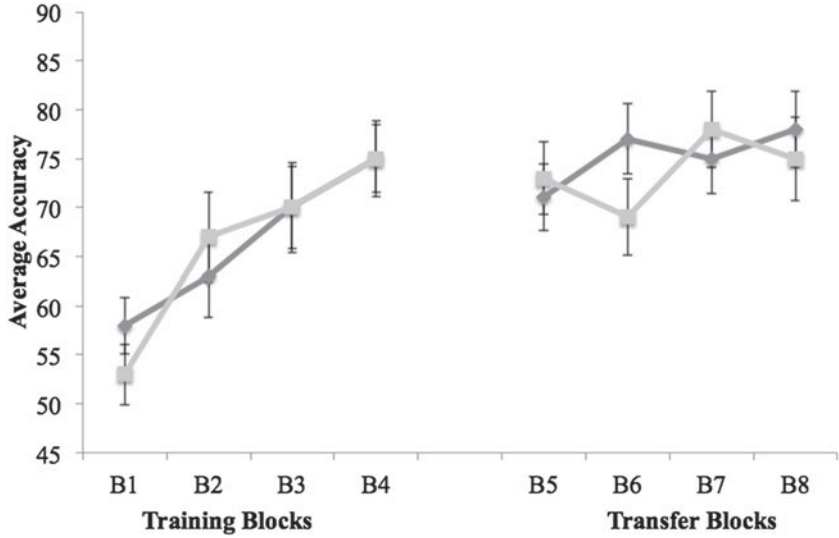


Fig. 4. Average accuracy by block in the labeled and unlabeled conditions for the training and transfer phases of the reversal shift task. Error bars represent Standard Errors.

### 5.3. DISCUSSION

Learning the original categorization with verbal labels did not reduce conceptual flexibility, as might be predicted by a perspective in which the top-down influence of verbal labels is strong and immutable at the level of representation connecting category and response. Instead, the reversal of responses to categories learned with verbal labels was relearned more quickly than the reversal of responses for categories that were not learned with verbal labels. While the cost of transfer was similar across label conditions, participants in the label condition demonstrated a quicker recovery, with improved accuracy in the second block after transfer.

This finding supports a view in which verbal label activate categories in such a way that their representation is more abstract in nature, with those features relevant to the category receiving top-down activation and those features irrelevant to the categorization receiving suppression. In such a way, labels can act as a stand-in for the mediating categorical representation between stimulus perception and overt response, with the ensuing on-line top-down modulation of lower levels of processing emphasizing category relevant features (Clark, 2013). This creates a category representation that is more abstract in that it predicts or fills in the stimulus-to-category associations. Flexibly adjusting behavior is then facilitated by the continued use of more abstract representation at relearning.

## 6. General discussion

The key result of these studies is the distinct influence of labels on recovery across the two different types of transfer. In the extra-dimensional switch task there was a high cost of transfer. These participants had to relearn their categorization strategies based on a previously unimportant dimension, at a steep cost and with slower relearning. Those participants who learned during training to categorize based on the spatial frequency of the eyes and had to switch during transfer to categorize based on the orientation demonstrated reduced ability to flexibly adjust to this new categorization strategy, a replication of prior transfer tasks (e.g., Goldstone & Steyvers, 2001). While selective attention is an important process in the development of accurate categorization (Goldstone, 1998), it also reduces the degree of flexibility present in responding in ways that cut across category boundaries. Importantly, however, the presence of labels in initial learning did not modulate the cost of transfer, or the relearning trajectory after transfer, providing convergent evidence that the on-line warping of selective attention in categorical perception is transient, and in this case weak enough not to impede relearning. Having learned the categories with labels originally, activation of the labels, and associated feedback, can be expected to have continued during the transfer phase of this condition. The top-down modulation from the labels activated in such a way, as opposed to overt activation through the presentation of the label, did not reduce recovery after transfer.

In contrast, for those in the reversal shift experiment, where participants had to learn that those aliens who had been approachable were now to be avoided and vice versa, there was a faster recovery after transfer when the original categories had been learned with labels during initial training than when they had been learned without labels. This suggests that labels play a positive role in the relearning of categorization when the relevant diagnostic dimension does not change, but the categorical behavioral responses to the two groups do. Having learned verbal labels for the categories allowed the participants to more flexibly adapt to the changing task demands. This finding is novel, and suggests, taken together with the lack of modulation of selective attention discussed above, that labels in categorization function as symbolic/material objects, anchoring thought and action (Clark, 2006, 2013).

Interestingly, the results of the present experiment did not find support for a general advantage for learning categories with labels over categories without labels, as seen in previous similar experiments (Lupyan et al., 2007). While numerical differences in accuracy across label type trended in this direction, this did not approach significance. It seems likely that the role of labels is strongest when the categories match historically predictive patterns, such as the shape-based aliens of the Lupyan et al. (2007) study. There are clear differences between the types of category. The stimuli used in Lupyan et al.

(2007) did not vary along two readily distinguishable dimensions, a requirement of the extra-dimensional shift task. While the finding is in line with previous research that indicates the label advantage may be restricted to shape-based categories (Brojde, Porter, & Colunga, 2011), similar gabor patch-based stimuli have been used previously for category learning, leading to an advantage for labeled categories of similar structure to those in the present study (Ketels & Jones, 2010). It is therefore unclear as to why the present experiment failed to replicate the positive effect of label on category learning. Given previous failures to replicate the finding, even with shape-based categories (Brojde et al., 2011), it seems likely that the modulation of perceptual category learning through words may simply be a weaker effect than has been suggested.

And indeed this weakness is an asset, as demonstrated by the present studies. If the top-down influence of verbal labels was too strong, or if the perceptual warping of the stimulus space more permanent, conceptual flexibility would be reduced. This finding suggests that words do not modulate categorical perception in such a way as to restrict behaviors to just those boundaries expressible in a lexicon. Learning categories with or without labels did not influence the cost of selective attention in switching between diagnostic dimensions of the stimulus set. In contrast with verbal interference studies that demonstrate the bleaching of cross-linguistic effects by removing language processing (Roberson & Davidoff, 2000; Winawer et al., 2007), the present study allowed for the continued activation of verbal labels post-transfer. While participants in both experiments reported in a post-study questionnaire that they continued to make use of the labels provided during training upon 'arriving on the planet', here the activation of categories through learned labels did not hinder relearning trajectories.

If labels had modulated the strength of selective attention in such a way as to restrict participants' ability to cross-cut learned category boundaries after training, we could then conclude that it would be possible for linguistic differences to lead to distinct patterns of non-verbal cognition. Instead, as has been demonstrated previously with object category labels (Malt, Slobin, & Gennari, 2003; Ross & Murphy, 1999), the effect of labels does not appear to influence the ability to think and act outside the encoding of one's particular language. While further tests of the role of labels on the ability to flexibly adjust the perceptual processes involved in categorization may be needed to fully illuminate the exact nature of the relationship between verbal and non-verbal thought, the findings presented here give support for a shallower interpretation of the Whorfian hypothesis.

These findings provide support for the conclusion that categories activated through labels are more abstract or 'categorical' in nature (Lupyan, 2008). One interpretation then is that a reversal shift is easier to learn when dealing

with a more abstract category, specifically when the category is learned with a verbal label. That flexibility is improved when representations are more abstract is supported by the active–latent account of cognitive flexibility (Kharitonova et al., 2009). Since verbal labels actively modulate conceptual representations of categories which they are used to express (Lupyan & Thompson- Schill, 2012), when the categories boundaries themselves do not change, but only the category-to-response associations, these labels continue to activate more abstract representations, simplifying the computation required and standing in as abstract symbols. It then becomes possibly a simpler task for the participants in the reversal shift transfer condition to switch from 'Gowachi' and 'Caleba' to 'not Gowachi' and 'not Caleba' compared to when the categories of aliens were learned without labels.

## 7. Conclusion

With habitual use of the specific set of conceptual symbolic representations afforded by a language, an individual may be biased towards these representations in problem-solving and other cognitive tasks. How a language may accomplish this is an arena of ongoing debate. The role that words play in categorization, and the degree to which words and their categories may be a unitary mental structure, is key to an understanding of both the general relationship between language and thought that has elevated us as a species, and may also provide evidence of the mechanism through which language-specific effects on thought have emerged. The label feedback hypothesis, as well as other theories, suggests that language augments cognition, with the activation of a word acting on the interactive system of categorization (Lupyan, 2012). This type of augmentation, when applied to a particular domain, allows for the distinctly human ability to flexibly interact with our environment, and exemplifies the way in which language provides a useful structure for thought (Clark, 2006; Vygotsky, 1986). Labels activate conceptual representations in a particularly effective way, with representations activated by words being more categorical in nature (Lupyan, 2008; Lupyan & Thompson-Schill, 2012).

Thus, the augmentation of perceptual processes involved in categorization appears to be just right, not too strong to reduce flexibility in re-categorizing the stimulus space, but strong enough to assist in flexibly changing response behaviors. Despite well-demonstrated influences of label activation directly on perception, here the verbal augmentation did not act to reduce flexibility in shifting attentional weighting from one stimulus dimension to another. Rather than modulating the mechanisms of categorization in such a way to reduce flexibility, words played a different role; that of 'mental anchors' for organizing and stabilizing thought (Clark, 2006). Given that words actually aided rather than reduced flexibility when the transfer task required a behavioral

rather than an attentional change, in tandem with the lack of effect on selective attention, suggests that the influence of language on thought is not as deep as some would suggest. To the extent that verbal labels play a role in shaping concepts as they are learned (Lupyan, et al., 2007; Lupyan & Thompson-Schill, 2012), the key factor appears to be the ability of words to stand in as category representations for thought. Rather than reducing the ability to think outside the structure provided by a particular language, languages provide us with the means to control and expand our repertoire of thoughts and behaviors.

## REFERENCES


Boroditsky, L., Schmidt, L. A., & Phillips, W. (2003). Sex, syntax, and semantics. In D. Gentner & S. Goldin-Meadow (Eds.), *Language in mind: advances in the study of language and thought* (pp. 61–80). Cambridge, MA: MIT Press.

Bowerman, M., & Choi, S. (2001). Shaping meaning for language: universal and language-specific in the acquisition of spatial semantic categories. In M. Bowerman & S. C. Levinson (Eds.), *Language acquisition and conceptual development* (pp. 475–522). Cambridge: Cambridge University Press.

Brojde, C. L., Porter, C., & Colunga, E. (2011). Words can slow down category learning. *Psychological Bulleting & Review*, **18**, 798–804.

Casasola, M. (2005). Can language do the driving? The effect of linguistic input on infants' categorization of support spatial relations. *Developmental Psychology*, **41**, 183–192.

Clark, A. (2006). Language, embodiment, and the cognitive niche. *TRENDS in Cognitive Science*, **10**, 370–374.

Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences*, **36**, 181–204.

Cohen, J. D., & Servan-Schreiber, D. (1992). Context, cortex, and dopamine: a connectionist approach to behavior and biology in schizophrenia. *Psychological Review*, **99**, 45–77.

Deacon, T. W. (1997). *The symbolic species: the co-evolution of language and the brain*. New York, NY: W.W. Norton.

DeLoache, J. S. (2004). Becoming symbol-minded. *Trends in Cognitive Science*, **8**, 66–70.

Gentner, D., & Goldin-Meadow, S. (2003). *Language in mind: advances in the study of language and thought*. Cambridge, MA: MIT Press.

Gilbert, A. L., Regier, T., Kay, P., & Ivry, R. B. (2006). Whorf hypothesis is supported in the right visual field but not the left. *Proceedings of the National Academy of Sciences*, **103**, 489–494.

Gleitman, L., & Papafragou, A. (2005). Language and thought. In K. Holyoak & B. Morrison (Eds.), *Cambridge handbook of thinking and reasoning* (pp. 633–661). Cambridge: Cambridge University Press.

Goldstone, R. L. (1994). Influences of categorization on perceptual discrimination. *Journal of Experimental Psychology: General*, **123**, 178–200.

Goldstone, R. L. (1998). Perceptual learning. *Annual Review of Psychology*, **49**, 585–612.

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

Goldstone, R. L., & Steyvers, M. (2001). The sensitization and differentiation of dimensions during category learning. *Journal of Experimental Psychology: General*, **130**, 116–139.

Gumperz, J. J., & Levinson, S. C. (1996). *Rethinking linguistic relativity*. Cambridge: Cambridge University Press.

Haider, H., & Frensch, P. A. (1996). The role of information reduction in skill acquisition. *Cognitive Psychology*, **30**, 304–337.

Harnad, S. (2005). Cognition is categorization. In H. Cohen & C. Lefebvre (Eds.), *Handbook of categorization in cognitive science* (pp. 19–43). San Diego: Elsevier.

Hendrickson, A. T., Kachergis, G., Fausey, C. M., & Goldstone, R. L. (2012). Re-learning labeled categories reveals structured representations. Paper presented at the Proceedings of the 34th annual meeting of the Cognitive Science Society, Sapporo, Japan.
Hespos, S., & Spelke, E. (2004). Conceptual precursors to language. *Nature*, **430**, 453–455.
Honey, R. C., & Hall, G. (1989). The acquired equivalence and distinctiveness of cues. *Journal of Experimental Psychology: Animal Behavior Processes*, **15**, 338–346.
Kay, P., & Kempton, W. (1984). What is the Sapir–Whorf hypothesis? *American Anthropologist*, **86**, 65–79.
Kendler, H. H., & Kendler, T. S. (1962). Vertical and horizontal processes in problem solving. *Psychological Review*, **69**, 1–16.
Ketels, S. L., & Jones, M. (2010). Verbal labels are not always useful. Poster presented at the Ninth Annual Summer Interdisciplinary Conference, Bend, OR.
Kharitonova, M., Chien, S., Colunga, E., & Munakata, Y. (2009). More than a matter of getting 'unstuck': flexible thinkers use more abstract representations than perseverators. *Development Science*, **12**, 662–669.
Kruschke, J. K. (1996). Dimensional relevance shifts in category learning. *Connection Science*, **8**, 201–223.
Levinson, S. C. (1997). From outer to inner space: linguistic categories and non- linguistic thinking. In J Nuyts & E. Pederson (Eds.), *Language and conceptualization* (pp. 13–45). Cambridge: Cambridge University Press.
Li, P., & Gleitman, L. (2002). Turning the tables: language and spatial reasoning. *Cognition*, **83**, 265–294.
Lupyan, G. (2007). The label feedback hypothesis: linguistic influences on visual processing. Unpublished doctoral dissertation, Carnegie Mellon University, Pittsburgh, Pennsylvania.
Lupyan, G. (2008). From chair to 'chair': a representational shift account of object labeling effects on memory. *Journal of Experimental Psychology: General*, **137**, 348–369.
Lupyan, G. (2009). Extracommunicative functions of language: verbal interference causes selective categorization impairments. *Psychonomic Bulletin & Review*, **16**, 711–718.
Lupyan, G. (2012). Linguistically modulated perception and cognition: the label-feedback hypothesis. *Frontiers in Psychology*, **3**, 1–13.
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.
Lupyan, G., & Spivey, M. (2010). Redundant spoken labels facilitate perceptions of multiple items. *Attention, Perception, & Psychophysics*, **72**, 2236–2253.
Lupyan, G., & Thompson-Schill, S. L. (2012). The evocative power of words: activation concepts by verbal and nonverbal means. *Journal of Experimental Psychology: General*, **141**, 141–170.
Maddox, W. T., Glass, B. D., O'Brien, J. B., Filoteo, J. V., & Ashby, F. G. (2010). Category label and response location shifts in category learning. *Psychological Research*, **74**, 219–236.
Malt, B. C., Sloman, S. A., & Gennari, S. (2003). Universality and language specificity in object naming. *Journal of Memory and Language*, **49**, 20–42.
Munakata, Y. (1998). Infant perseveration and implications for object permanence theories: a PDP model of the A-not-B task. *Developmental Science*, **1**, 161–211.
Nosofsky, R. M. (1986). Attention, similarity, and the identification–categorization relationship. *Journal of Experimental Psychology: General*, **115**, 39–57.
Pinker, S. (1995). *The language instinct: the new science of language and mind*. New York, NY: Penguin Books.
Roberson, D., & Davidoff, J. (2000). The categorical perception of colors and facial expressions: the effect of verbal interference. *Memory and Cognition*, **28**, 977–986.
Roberson, D., Davidoff, J., Davies, I. R. L., & Shapiro, L. R. (2005). Color categories: evidence for the cultural relativity hypothesis. *Cognitive Psychology*, **50**, 378–411.
Ross, B. H., & Murphy, G. L. (1999). Food for thought: cross-classification and category organization in a complex real-world domain. *Cognitive Psychology*, **38**, 495–553.
Shiffrin, R. M., & Schneider, W. (1977). Controlled and automatic human information processing: II. Perceptual learning, automatic attending and a general theory. *Psychological Review*, **84**, 127–190.

Spelke, E. S., & Tsivkin, S. (2001). Initial knowledge of conceptual change: space and number. In M. Bowerman & S. C. Levinson (Eds.), *Language acquisition and conceptual development* (pp. 475–511). Cambridge: Cambridge University Press.

Vygotsky, L. (1986). *Thought and language* (rev. ed.). Cambridge, MA: MIT Press.

Waxman, S. R., & Markow, D. B. (1995). Words as invitation to form categories: evidence from 12- to 13-month-old infants. *Cognitive Psychology*, **29**, 257–302.

Winawer, J., Witthoft, N., Frank, M., Wu, L., Wade, A., & Boroditsky, L. (2007). Russian blues reveal effects of language on color discrimination. *Proceedings of the National Academy of Sciences*, **104**, 7780–7785.

Yoshida, H., & Smith, L. B. (2005). Linguistic cues enhance the learning of perceptual cues. *Psychological Science*, **16**, 564–577.