Verbal labels promote representational alignment even in the absence of communication

Ellise Suffill (suffill@wisc.edu)
Department of Psychology, University of Wisconsin-Madison,
1202 W. Johnson Street, Madison, WI 53706 USA

Gary Lupyan (lupyan@wisc.edu)
Department of Psychology, University of Wisconsin-Madison,
1202 W. Johnson Street, Madison, WI 53706 USA

Abstract

What affects whether one person represents an item in a similar way to another person? We examined the role of verbal labels in promoting representational alignment. Three groups of participants sorted novel shapes on perceived similarity. Prior to sorting, participants in two of the groups were pre-exposed to the shapes using a simple visual matching task and in one of these groups, shapes were accompanied by one of two novel category labels. Exposure with labels led people to represent the shapes in a more categorical way and to increased alignment between sorters, despite the two categories being visually distinct and participants in both pre-exposure conditions receiving identical visual experience of the shapes. Results hint that labels play a role in aligning people's mental representations, even in the absence of communication.

Keywords: categorization; labels; alignment; coherence.

Introduction

How similar is one person's representation of an item to that of other people's? The same items can be represented in many different ways and, as such, categories of items can also vary. However, in communication between individuals we need to align upon how we represent items and categories of items if we are to successfully communicate about things in the world (Markman & Makin, 1998; Pickering & Garrod, 2004; Silvey, Kirby & Smith, 2019). Given the wide variability in how we can represent items, how is it that our representations align?

Past research suggests that labels can promote the alignment of categories both with communication about category items (Markman & Makin, 1998), and without communication about category items (Suffill, Branigan & Pickering, 2016; 2019). But by what mechanism do labels increase category alignment across people? The label-feedback hypothesis (Lupyan, 2012) predicts that perceptual input that has been previously associated with a label will automatically activate the label and the label will in turn selectively activate category-diagnostic features, causing the representation of the item to become more categorical.

For example, after hearing basic color names such as "red" and "blue" people are more accurate in discriminating category members from non-members, and in discriminating typical members from atypical ones. While hearing a categorical color label affects discrimination, seeing a visual cue (e.g., a specific shade of red) does not (Forder & Lupyan, 2019). Labels have also been found to influence visual search for numbers and objects (Lupyan & Spivey, 2010; Gilbert et al., 2008), perception of orientation (Smilek, Dixon & Merikle, 2006), and facial expressions (Roberson & Davidoff, 2000; Brook et al., 2016). In all these cases, the labels appear to induce a more categorical representation, in particular, a representation that emphasizes category diagnostic features of the named category – the features that most reliably distinguish category members from non-members.

One consequence of this increased categoricality may be greater alignment between people. For example, when a category label ("triangle") is used as a cue, people appear to activate more typical equilateral triangles (Lupyan, 2015) compared to when they are cued by definitionally equivalent cues like "three-sided polygon". Equilateral triangles are more similar to one another than those judged as less typical (e.g., scalene). And so, to the extent that "triangle" causes people to think about an equilateral triangle, the category label is, in effect, aligning people's representations. Under the influence of the label, people's representations of "a triangle" are thus more similar than they would be otherwise.

In previous work, Suffill et al. (2016; 2019) have shown that when asked to group items into categories labeled with nonsense labels, people produce more similar groups than when asked to group the same items into unlabeled categories. Here, we test the hypothesis that exposure to labels promotes alignment in a more systematic way.

Current study

To examine whether labels promoted greater alignment between people by increasing the categoricality of their representations, we familiarized people with two visually distinct categories of novel shapes with the categories either

labeled or not, and then probed their representational similarity of the shapes using a sorting task (Goldstone, 1994; Malt, Sloman, Gennari, Shi & Wang, 1999). We predicted that although the structure of the categories made it plain that there were two distinct kinds, exposure to labels would cause people to represent the items in a more categorical way (i.e., emphasizing the category-diagnostic features if the shapes) and, as a result, would tend to help people align to a greater extent.

Method

Participants

We recruited 129 (85 female) Psychology students at the University of Wisconsin-Madison, who took part for course credit. Ages: 18-22 years ($\bar{x}$ = 18.77, *SD* = 0.68). Participants were randomly assigned to the "Baseline" (*N* = 45), "No Labels" (*N* = 43) or "With Labels" (*N* = 41) conditions[1].

Stimuli

We generated a prototype shape for each of the two categories (i.e., generically named category "A" and "B"; Fig. 1). Our aim was to create two categories that were visually distinct but for which these distinctions were difficult to label, in order to avoid simple linguistic distinctions like 'smooth' versus 'pointy'. To create category members with a family resemblance structure, we generated distortions by adding varying amounts of random gaussian noise to the coordinates of the prototypes. We generated an additional 18 shapes per category by adding noise at three different thresholds to produce category members that were a "low" (*N* = 6; $\bar{x}$ *distance* = .21), "medium" (*N* = 6; $\bar{x}$ *distance* = .30) or "high" (*N* = 6; $\bar{x}$ *distance)* = .40) level of distortion from their prototype (as measured by Euclidean distance). This resulted in 19 shapes (including the prototype) per category. The labels were two nonsense words ("talp" and "gek") recorded by an American English speaker. To equate auditory exposure, participants in the "No Labels" condition heard length and volume-normalized white noise in place of the labels.

Procedure

Pre-exposure. Participants assigned to the "With Labels" or "No Labels" conditions began by completing a delayed match-to-sample task that served to familiarize people with the visual stimuli and, for the "With Labels" condition, expose people to the labels (Fig. 2). On each trial, participants saw one of the shapes (sample) which was either labeled or not depending on condition. After a delay, participants saw two shapes and had to indicate which one matched the sample. There was a total of 243 trials (3 blocks of 9 shapes from each of the two categories, paired with 9 shapes from the other category. Category prototypes were omitted. Importantly, the two shapes presented side-by-side were always from different categories, thus giving participants from the "With Labels" and "No Labels" conditions equal experience with making between-category discriminations. The display remained visible until a response was made. The correct response was counterbalanced across left and right positions. Errors were signaled with a buzzing sound.

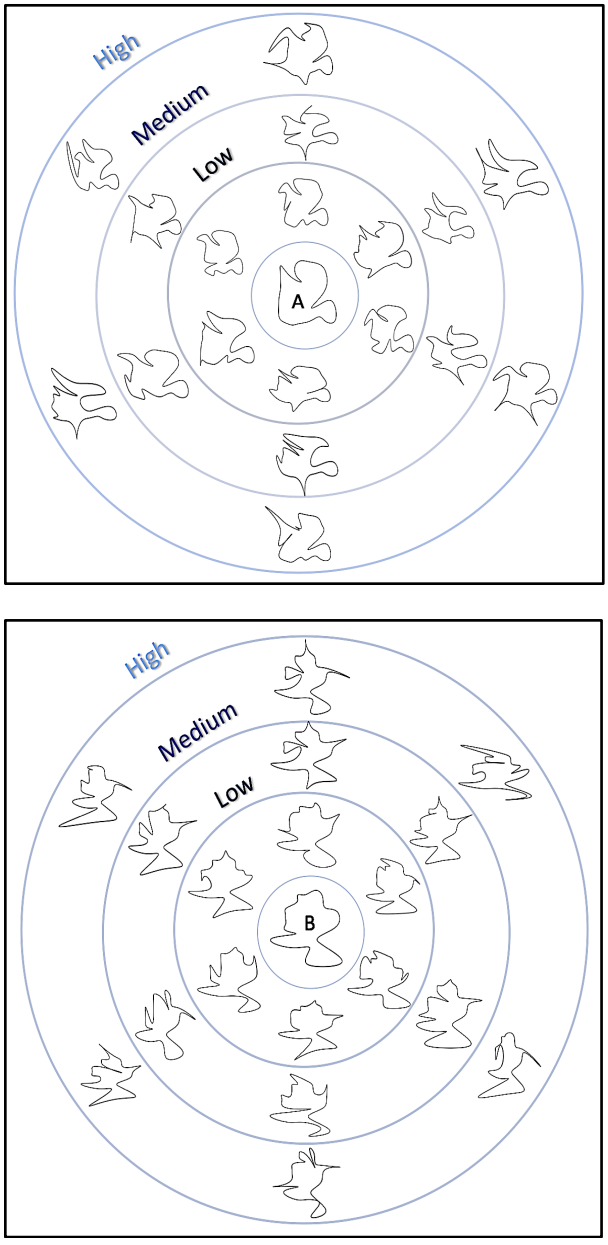

Figure 1: Category A (Top) and B (Bottom) prototypes with "low", "medium" and "high" distortion. Shape position within sections is random.

[1] We excluded 39 participants who did not move items during the free sort as analyses required all items to be meaningfully sorted by perceived similarity (i.e., instead of being left in random starting positions). We subsequently modified the instructions to emphasize that all items had to be moved.

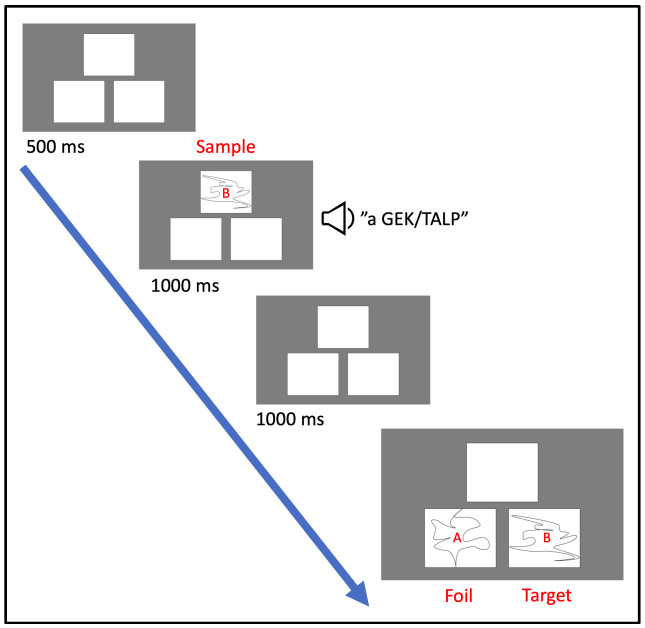

Figure 2: Schematic of delayed match-to-sample task used for pre-exposure ("With Labels" condition). Participants responded with which shape matched the sample. A/B category labels are shown for illustration only.

Free sort. Participants were presented with 20 shapes (i.e., 10 A category shapes; 10 B category shapes), including the previously unseen prototype shape, six further novel shapes and three previously seen shapes for each category. Shapes were initially displayed around the four edges of the screen, and participants were asked to drag the shapes into any number of categories on the basis of similarity. We did not provide predefined spaces for the formation of categories: Instead, participants were able to place the shapes into any number of categories by spatially clustering the shapes together. Shapes were allowed to overlap. Participants were instructed to move all of the shapes during sorting, in order to ensure that the final positions of shapes were meaningful to each sort (i.e., so that the final position of a shape was not simply its random starting location). We also tested 7 of the participants in the "With Labels" condition for label retention, i.e., whether they accurately remembered which shapes were 'talps' and which were 'geks' following the Free sort phase. Average accuracy was .87 (*SD* = .19), suggesting that participants tended to remember the labels despite their incidental nature.

Results

Analytic Approach

For Pre-exposure, we examined differences in accuracy and reaction time in the match-to-sample task for the "No Labels" and "With Labels" conditions. For the Free sort, we first assessed the average Euclidean distances for between- and within-category items to check whether participants across all conditions were sensitive to the visual differences across the categories ("Within versus between category distances"). We then assessed how participants sorted the items across conditions: we examined the tendency for participants to use different numbers of clusters in their solutions ("Number of clusters"); the properties of the clusters ("Cluster properties"); and finally how similar participants' sorts were across participants ("Effects of labels on alignment").

We used mixed effects linear models for continuous output variables and logistic regression for discrete variables, as implemented in R's lme4 package v. 1.1-21 (Bates et al., 2015). Predictors were center-coded. Models included by-subject random intercepts and random-slopes for within-condition factors unless doing so prevented convergence. All reported models were a significantly better fit of the data than null models ($p < .05$).

Pre-exposure phase

Average accuracy. Average accuracy on the delayed match-to-sample task was $\bar{x} = 0.98$ ($SD = 0.13$) for the "No Labels" condition nearly identical, $\bar{x} = 0.98$ ($SD = 0.14$) for the "With Labels" condition. There was no significant difference in accuracy between any of the conditions ($p = .81$).

Average reaction time. Average reaction times (trimmed to exclude RTs > 2 *SD* from the overall mean) for correct responses were marginally faster for the "No Labels" condition ($\bar{x} = 648$ ms, $SD = 543$ ms) compared to the "With Labels" condition ($\bar{x} = 731$ ms, $SD = 1736$ ms) ($b = -54.13$, $SE = 27.70$, $t = -1.95$, $p = .05$).

Free sort phase

Next, we examined how people subsequently sorted new and previously experienced shapes, including the category prototypes (see Fig. 3 for an example sorting solution).

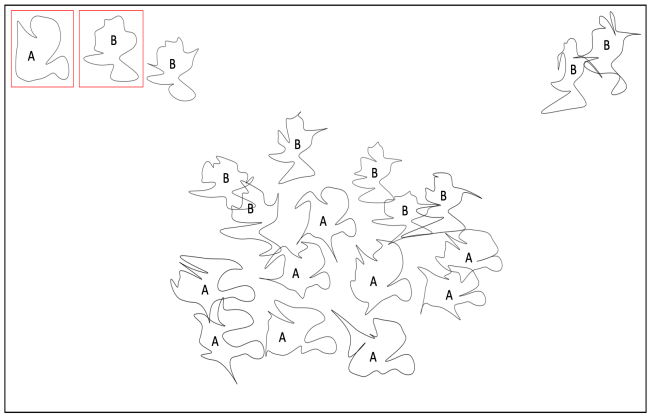

Figure 3: Example of a sorting solution from a participant in the "Baseline" condition. Category identity (A/B) and prototypes (in red) are for illustration only.

Within versus between category distances. We computed the Euclidean distance (in pixels) between each pair of sorted shapes within-category (e.g., every A1-A2, A2-A3, B1-B2) and compared the mean distances to between-category pairs (A1-B1, A1-B2, etc.).

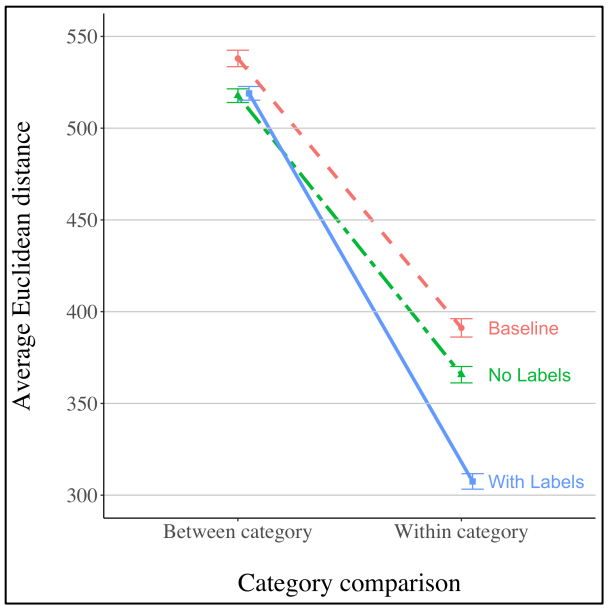

Figure 4: Average Euclidean distance (in pixels) between pairs of shapes that span a category boundary (between-category) vs. not (within-category). Error bars denote the standard error of the mean.

Figure 4 shows a strong main effect of comparison-type: all groups placed between-category items farther apart than within-category items (b = -169.11, SE = 13.55, t = -12.48, p < .001). That participants in the "Baseline" condition were able to make the within vs. between-category distinction demonstrates that people could distinguish the two categories even without pre-exposure. The difference was significantly more pronounced in the "With Labels" condition compared to both the "Baseline" (b = -64.65, SE = 7.75, t = -8.35, p < .001) and "No Labels" (b = -59.44, SE = 7.43, t = -8.00, p < .001) conditions. The "Baseline" and "No Labels" conditions did not significantly differ from one another in the difference between within- and between-category distances (p = .51). In sum, the results show that labels lead to more categorical item placement while pre-exposure on its own does not.

Number of clusters. We used the "pamk" function ("fpc" package; Hennig, 2019) to group each person's final item locations into medoid-based clusters[2]. Participants in the "With Labels" condition ($\bar{x}$ = 3.10, SD = 1.60) formed significantly fewer clusters than participants in the "No Labels" ($\bar{x}$ = 4.07, SD = 2.00) (t(79.47) = 2.47, p = .02) and "Baseline" ($\bar{x}$ = 3.93, SD = 1.50) conditions (t(82.01) = 2.50, p = .01). The number of clusters in the "Baseline" condition did not significantly differ from the "No Labels" condition (p = .72). We also assessed how likely participants were to use two clusters. "With Labels" participants were significantly more likely to use a 2-category solution (N = 22/41), compared to participants in the "Baseline" condition (N = 9/45) (X^2(1) = 9.13, p = .003). There was no significant difference between the "With Labels" and "No Labels" (N = 14/43) conditions (p = .08), or between the "No Labels" and "Baseline" conditions (p = .27).

Cluster properties. We next examined the kinds of items participants clustered together. The first property we examined was cluster *purity*. A cluster had a purity of 1 if all the shapes were from the same category and a purity of .50 if it contained an equal number of A and B category shapes. Because cluster purity is inversely correlated with the number of items in a cluster, we used a weighted regression where purity was weighed by cluster size. There were no differences in purity between "Baseline" ($\bar{x}$ = .86, SD = .17) and "No Labels" ($\bar{x}$ = .90, SD = .16) (p = .28) or "With Labels" ($\bar{x}$ = .88, SD = .17) (p = .31). There was also no significant difference in purity between "No Labels" and "With Labels" (p = .96). We next examined purity more selectively (see Fig. 5)[3]: looking specifically at the clusters that contained the A or B prototype. This revealed that clusters containing the A or B prototypes had greater purity in the "With Labels" condition, than the "Baseline" (b = 0.73, SE = 0.20, t = 3.57, p < .001) and "No Labels" (b = 1.10, SE = 0.43, t = 2.55, p = .01) conditions (see within-category vs. between-category comparison in Fig. 6). That is, participants in the "With Labels" condition were more likely to put A items in a cluster containing the A prototype (and vice versa), than were participants in the other conditions. There was no significant difference in prototype cluster purity between the "Baseline" and "No Labels" conditions (p = .25).

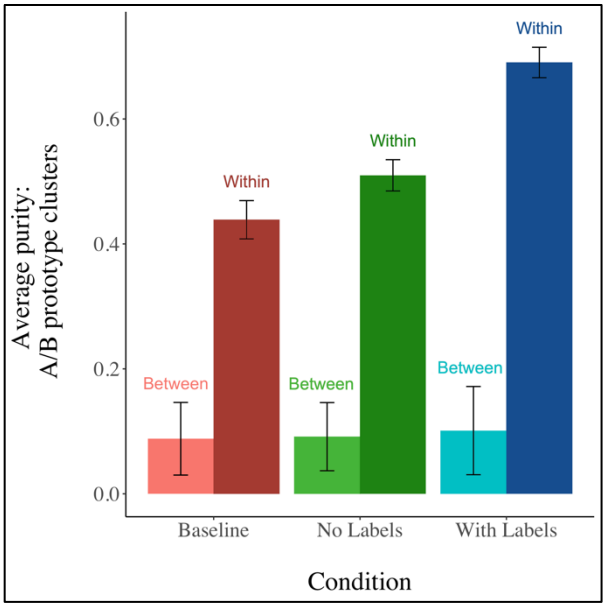

Figure 5: Composition of clusters containing prototypes for "Within" (e.g., A item + A prototype) and "Between" (e.g., A item + B prototype) category comparisons. Error bars denote the standard error of the mean.

[2] A medoid is the item within a cluster for which the average distance between it and all other cluster members is smallest (Kaufman & Rousseeuw, 1990).

[3] For this analysis, we removed data that clustered both prototypes into one cluster. Adjusted N: "Baseline": 29/45; "No Labels": 34/43; "With Labels": 28/41.

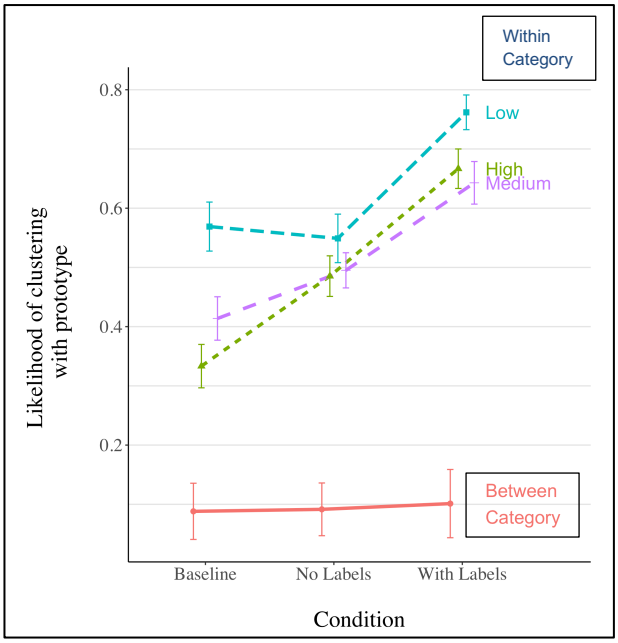

Figure 6: Average likelihood of clustering within-category "low", "medium" and "high" distortion items (i.e., A-A; B-B) versus all between-category items (A-B) with the prototypes. Error bars denote the standard error of the mean.

We also examined how likely participants were to sort "low", "medium" and "high" distortion items into the same clusters as the prototype (see clustering of "low" vs. "medium" vs. "high" distortion items for within-category comparisons in Fig. 6). There was a significant effect of Distortion, such that participants were less likely to cluster more distorted items with the prototype, compared with less distorted items (b = -0.35, SE = 0.07, t = -4.82, p < .001).

Effects of labels on alignment. Having shown that exposure to labels results in more categorical sorting solutions, we can now ask whether labels also led people to form more similar sorts (i.e., whether labels led to greater alignment). We coded whether participants put each possible pair of shapes (20 x 19/2 = 190 shape pairs) into the same cluster. If a participant placed two shapes into the same cluster, that shape pair was coded as 1; if not, it was coded as 0. We then compared each pair of participants within a condition on how often they matched in categorizing shape pairs (i.e., if they were both assigned a 1 for a shape pair, they both received a match for that shape pair)[4]. We repeated this for all shape pairs, and used this to compute a proportional score of alignment for each participant pair (e.g., if a pair of participants matched on all 190 shape pairs, they would receive an alignment score of 1; if they matched on 50 shape pairs they would receive an alignment score of 50/190 = 0.26) (see Fig. 7 for average alignment scores by participant and condition). We took every participant from the "Baseline" condition and compared their data to every other participant from the "Baseline" condition to get each pair's alignment as a proportion; we repeated this process for the "No Labels" and "With Labels" conditions separately. Average alignment across the conditions was $\bar{x}$ = 0.10 (SD = 0.06) for the "Baseline" condition, $\bar{x}$ = 0.13 (SD = 0.09) for the "No Labels" condition, and $\bar{x}$ = 0.20 (SD = 0.11) for the "With Labels" condition.

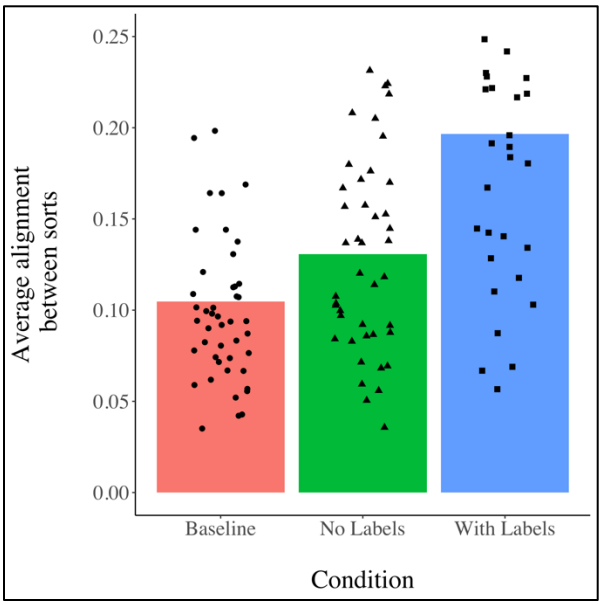

Figure 7: Average of alignment by Condition. Points represent the average of each participant's alignment score with every other participant within the condition.

We took the alignment scores for every possible participant pair within the three conditions and analyzed this by Condition with random intercepts by participant for each participant pair (i.e., arbitrarily participant 1 and participant 2). As the alignment score is proportional, we log-transformed alignment scores. The number of clusters each participant used can affect the chance alignment between a pair of participants (e.g., if participants in a pair both formed two clusters, their chance alignment would be .25; but if one participant used two clusters and the other used three, their chance alignment would be .17). To ensure that any condition differences in alignment did not simply reflect condition differences in cluster number, we statistically controlled for chance-level performance for each unique participant pairing. There were effects of Condition: Alignment was significantly higher in the "With Labels" condition compared to the "Baseline" condition (b = 0.45, SE = 0.11, t = 4.20, p < .001). And, more importantly, alignment was significantly higher in the "With Labels" condition, than in the "No Labels" condition (b = 0.31, SE = 0.11, t = 2.80, p = .01). The

[4] While this measure is similar to the Rand Index (Rand, 1971), alignment between two participants increases only when both place a given item pair in the same cluster, but not when both place an item pair into different clusters.

"Baseline" and "No Labels" conditions did not significantly differ in alignment ($p > .05$).

Discussion

We examined whether novel labels promote greater alignment between people by increasing the categoricality of their representations. Despite the two categories being highly discriminable (because they were generated by perturbing two rather differently shaped prototype shapes) and despite people having had the same amount of visual exposure to the categories, those who experienced the shapes alongside redundant non-word labels had more categorical representations of both novel and previously experienced shapes. Those who were exposed to labels during the pre-exposure phase (the "With Labels" condition) clustered items from the same category closer to one another and were more likely to group category prototypes with items from the same category. And, critically, these participants were more aligned with one another as demonstrated by more similar sorting solutions, than participants from the other two conditions.

Including the "Baseline" condition along with the two pre-exposure conditions allowed us to compare the effect of the presence of incidental labels while keeping visual and categorization experience constant – that is, the contrast between the "No Labels" vs. "With Labels" conditions – to the effect of the pre-exposure phase (243 trials of a delayed match-to-sample task) – that is, the contrast between the "Baseline" and "No Labels" conditions. The data show that for nearly all the analyses, it is the presence of labels that makes the larger difference to categoricality and alignment than the pre-exposure phase.

Together, these findings suggest that participants who were exposed to the shapes with labels produced more categorical representations of the shapes than did participants who received identical visual exposure to the category structure. We suggest that the informationally redundant novel labels caused people to form more categorical representations (Lupyan, 2012). Crucially, the category-diagnostic features in these representations are those most likely to be sensible to the majority of people (Suffill et al., 2019). The selection of category-diagnostic features subsequently results in greater alignment, compared with participants who received equal visual experience with the categories, but for whom the shapes remained unlabeled.

Our use of sorting as a way of measuring representational similarity has some notable limitations. Although it allows us to measure the similarity in cluster composition between people, it does not reveal the internal structure of each cluster (as intended by the sorter). And although we measure distance between items as analogous to the similarity of the items as perceived by the sorter, there is individual variation in whether participants treat item distance as a continuous measure of similarity or just arrange items into discrete "clumps" (Goldstone, 1994). One way to overcome these limitations in future work may be to emphasize that distance between both items and clusters should correspond to perceived similarity and to ask participants to place the item that is most characteristic of each cluster centrally within the cluster.

Our results show that even when people's perceptual experience is equated (as it is in the "No labels" and "With Labels" conditions), brief and incidental exposure to novel category labels can promote more categorical representations as evidenced by the larger separation between A and B category items in the Free sort solutions and the greater likelihood of grouping items with their category prototypes. Labels also promoted greater alignment as evidenced by "With Labels" participants having more similar Free sort solutions to one another, than participants in the "No Labels" or "Baseline" conditions. Even when people's perceptual experiences are equated, exposure to novel category labels appears to make people's representations more calibrated.

Acknowledgments

This research was supported by NSF BCS 1734260 to Gary Lupyan.

References

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. Journal of Statistical Software, 67(1), 1-48.

Brooks, J. A., Shablack, H., Gendron, M., Satpute, A. B., Parrish, M. H., & Lindquist, K. A. (2017). The role of language in the experience and perception of emotion: A neuroimaging meta-analysis. *Social Cognitive and Affective Neuroscience*, *12*(2), 169-183.

Forder, L., & Lupyan, G. (2019). Hearing words changes color perception: Facilitation of color discrimination by verbal and visual cues. Journal of Experimental Psychology: General, 148(7), 1105.

Gilbert, A. L., Regier, T., Kay, P., & Ivry, R. B. (2008). Support for lateralization of the Whorfian effect beyond the realm of color discrimination. Brain & Language, 105, 91-98.

Goldstone, R. (1994). An efficient method for obtaining similarity data. *Behavior Research Methods, Instruments, & Computers*, *26*(4), 381-386.

Hennig, C. (2019). fpc: Flexible procedures for clustering. R package version 2.2-3.

Kaufman, L., & Rousseeuw, P. J. (1990). *Finding groups in data: an introduction to cluster analysis*. Hoboken, NJ: John Wiley & Sons.

Lupyan, G. (2012). Linguistically modulated perception and cognition: the label-feedback hypothesis. *Frontiers in Psychology, 3*, 54.

Lupyan, G. (2015). The paradox of the universal triangle: Concepts, language, and prototypes. *The Quarterly Journal of Experimental Psychology*, *70*(3), 389-412.

Lupyan, G., & Spivey, M. J. (2010). Redundant spoken labels facilitate perception of multiple items. Attention, Perception, & Psychophysics, 72(8), 2236-2253.

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*(12), 1077-1083.

Malt, B. C., Sloman, S. A., Gennari, S., Shi, M., & Wang, Y. (1999). Knowing versus naming: Similarity and the linguistic categorization of artifacts. *Journal of Memory and Language*, *40*(2), 230-262.

Markman, A. B., & Makin, V. S. (1998). Referential communication and category acquisition. Journal of Experimental Psychology: General, 127(4), 331-354.

Peirce, J. W., Gray, J. R., Simpson, S., MacAskill, M. R., Höchenberger, R., Sogo, H., Kastman, E., & Lindeløv, J. (2019). PsychoPy2: experiments in behavior made easy. Behavior Research Methods.

Pickering, M. J., & Garrod, S. (2004). Toward a mechanistic psychology of dialogue. Behavioral and Brain Sciences, 27(2), 169-190.

Rand, W. M. (1971). Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association, 66*(336), 846–850.

R Development Core Team. (2019). R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing.

Roberson, D., & Davidoff, J. (2000). The categorical perception of colors and facial expressions: The effect of verbal interference. Memory & Cognition, 28, 977-986.

Silvey, C., Kirby, S., & Smith, K. (2019). Communication increases category structure and alignment only when combined with cultural transmission. Journal of Memory and Language, 109, 104051.

Smilek, D., Dixon, M. J., & Merikle, P. M. (2006). Revisiting the category effect: The influence of meaning and search strategy on the efficiency of visual search. Brain Research, 1080, 73-90.

Suffill, E., Branigan, H. P., and Pickering, M. J. (2016). When the words don't matter: arbitrary labels improve categorical alignment through the anchoring of categories. In A. Papafragou, D. Grodner, D. Mirman, & J. C. Trueswell (Eds.). Proceedings of the 38th Annual Conference of the Cognitive Science Society. Austin, TX: Cognitive Science Society.

Suffill, E., Branigan, H., & Pickering, M. (2019). Novel labels increase category coherence, but only when people have the goal to coordinate. Cognitive Science, 43(11), e12796.