Running Head: CATEGORY LABELS

Immediate and Sustained Effects of Verbal Labels for Newly-Learned Categories

Fotis A. Fotiadis

Department of Psychology, Panteion University of Social and Political Sciences, Greece

Athanassios Protopapas

Department of Special Needs Education, University of Oslo, Norway

In press, *Quarterly Journal of Experimental Psychology*

Address correspondence to:

Fotis A. Fotiadis

Department of Psychology

Panteion University of Social and Political Sciences

Leoforos Syngrou 136

GR-176 71 ATHINA

Tel: +30 964 7941995

E-mail: fotis.fotiadis@panteion.gr

Abstract

Labels for the categories have been found to facilitate learning by boosting accuracy. According to the label-feedback hypothesis this facilitation is due to a mechanism selectively sensitizing perceptual dimensions. To further investigate the label facilitation phenomenon, one group of participants in our study learned both named and hard-to-name artificial categories, in a novel, within-subjects design. Another group of participants was administered a—highly similar—paired-associate task purportedly not involving sensitization of dimensions. Results showed that labels boosted accuracy during learning, but only when learning to categorize—not when learning to associate. The label-feedback hypothesis posits that labels exert an influence also after new categories have been learned. To test for sustained effects of labels, we administered a post-learning visual discrimination task while monitoring participants' eye movements and analyzing dwell time on the trained shapes. There was some indication of sustained effects of labels for newly-learned categories, but there was no effect following learning to associate. Our results suggest that labels for newly-learned categories have immediate effects during learning, and that the effects of labels may also be sustained during post-learning processing.

Keywords: verbal labels, category learning, visual attention, eye tracking

Introduction

In the field of category learning, verbal labels for the categories have been argued to influence categorization processes. Evidence to support this link originates in developmental psychology, as children's formation of categories is affected by correlated linguistic cues (Landau et al., 1988; Yoshida & Smith, 2005). Children's categorization is facilitated when categories are accompanied by verbal labels (Waxman & Markow, 1995), a benefit not observed with other kinds of cues (such as tones; Fulkerson & Waxman, 2007).

The facilitative effect of verbal labels for the categories is also evident in adults with fully developed linguistic capacities. Lupyan et al. (2007) trained participants to classify figures of alien creatures in two categories. Following each categorization decision, a redundant verbal label was presented. Results suggested that verbal labels (either visual or auditory) for the categories facilitated learning by increasing accuracy during the learning process, compared to non-verbal (location) cues or to the absence of cues. Thus, verbal labels for the categories have been suggested to have immediate effects on learning to categorize.

The effects of labels are also thought to persist after the categories have been formed. Labels for the categories may interact with the perceptual processing of categorization for items in well-known categories ("concepts", e.g., Lupyan, 2008b; Lupyan & Thompson-Schill, 2012, for a review see Lupyan et al., 2020). Moreover, labels can exert an influence following the learning of novel artificial categories (Tolins and Colunga, 2015). Thus, the effects of verbal labels for the categories are not only immediate but also expected to be sustained.

This study re-examines the effects of category labels for newly-trained artificial categories. Both immediate and sustained effects were tested, using a novel within-subject

design consisting of two sessions, i.e., two successive tasks. For clarity purposes, we present the study of immediate and sustained effects separately, but the reader should keep in mind that participants took part in a single experimental session.

Study of Immediate Effects

The label-feedback hypothesis (Lupyan, 2012a; 2012b) was founded on work suggesting that categorization is accompanied by the warping of perceptual space (Goldstone, 1994; Goldstone & Steyvers, 2001). Specifically, Goldstone (1994) showed that perceptual dimensions that are category-relevant are sensitized, whereas category-irrelevant dimensions are desensitized, as evident by post-categorization perceptual judgments (see also Folstein et al., 2012, 2013, 2014, 2015; Pothos & Reppa, 2014; Van Gulick & Gauthier, 2014). Lupyan (2012a; 20212b) suggested that this warping of perceptual space—or, in other words, a selective activation of category-diagnostic perceptual features—is boosted when verbal labels for the categories are present (but not when labels are absent or when linguistic processes are down-regulated). The influence of labels results in more "prototypical" (Lupyan, 2012b) or "categorical" (Lupyan, 2012a) perceptual representations of categorization items, in the sense that perceptual differences that are important for categorization are emphasized whereas unimportant differences are deemphasized.

The mechanism of selective sensitization has also received support from a recent study by Barnhart et al. (2018) employing a familiarization paradigm. Children and young adults passively viewed items of a category accompanied by an auditory label, and eye recordings revealed that fixations on category-diagnostic features of categorization item increased in the course of the familiarization. A result of this purported sensitization mechanism, as predicted by Lupyan (2012b), is that the process of category learning is

facilitated (i.e., accuracy is increased) when labels for the categories are present (Lupyan et al., 2007).

However, the selective-sensitization mechanism has not always been found to boost categorization accuracy. Specifically, in category structures that may be learned using more than one dimension, an attentional shift has been revealed. In particular, labels were found to promote selective activation of perceptual dimensions that are typically diagnostic of category membership in real-world occasions, such as shape (over hue, Brojde et al., 2011) or frequency (over orientation, Perry & Lupyan, 2014), even if this shift had deleterious effects on accuracy within the specific experimental context. Thus, the original prediction has been refined by empirical evidence: Labels may be predicted to boost learning accuracy, but only when the category-relevant perceptual dimension is typically diagnostic in everyday categorization.

Nevertheless, a review of the literature regarding the learning of such "everyday" categories suggests that the finding of label facilitation is not ubiquitous. Brojde et al. (2011) used the stimuli of Lupyan et al. (2007), and found in two experiments that labels had no effect on shape-based categorization accuracy. Using a similar procedure, Tolins and Colunga (2015) found no label advantage during learning to categorize when frequency was the category-diagnostic dimension. Finally, Lupyan and Casasanto (2015) contrasted the effects of different redundant verbal labels on shape-based categorization performance. They found an advantage from redundant real words, compared to the no-label condition, as well as from pseudowords, but only when they were selected to activate the same class of perceptual features as the words.

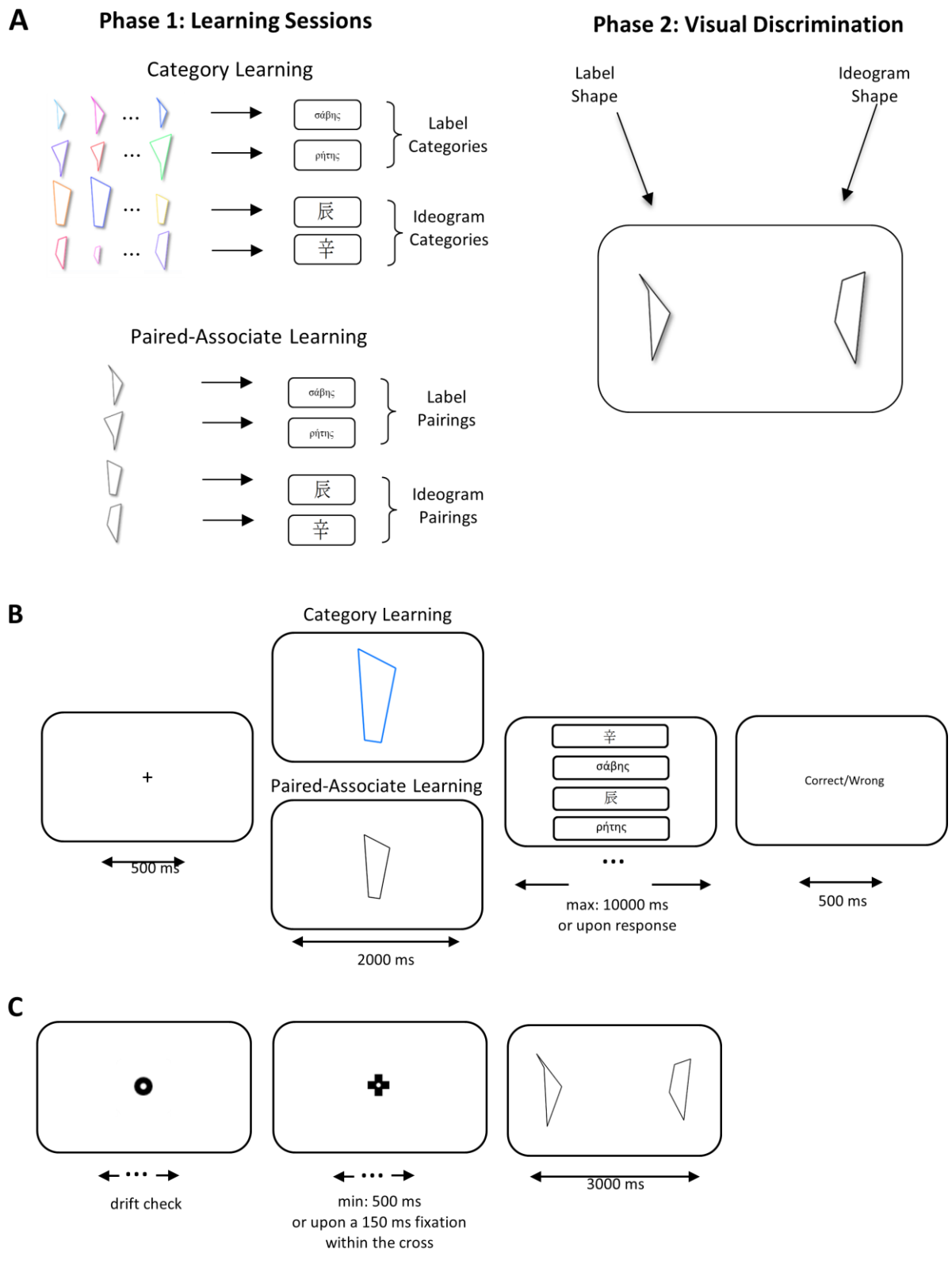
In sum, the facilitative effect of labels during the learning of categories based on dimensions that are typically predictive of category membership (Lupyan et al., 2007) has

occasionally failed to emerge (Brojde et al., 2011; Tolins & Colunga, 2015), or has been subject to choices regarding the experimental materials (Lupyan & Casasanto, 2015). This constitutes a challenge for the label-feedback hypothesis and warrants further examination of the replicability of the phenomenon. This was the purpose of the present study.

To test for the label-facilitation phenomenon during learning, we asked a group of participants to learn four novel artificial categories. Categorization stimuli varied across the dimensions of shape, color and size. Importantly, the category-relevant dimension was shape, a dimension that is typically predictive of categories. Instead of using redundant verbal labels after a categorization decision has been made (Lupyan et al., 2007), we tested for the effect of labels by manipulating the nameability of the category labels: Two of the categories were denoted by pseudowords (label categories), and two of the categories were denoted by visual symbols (ideogram categories, see Fig. 1A) that were previously found to be hard to name (Fotiadis & Protopapas, 2014). We predicted that participants should learn the label categories with increased accuracy compared to the ideogram categories.

Figure 1

(A) Design of the Experiment and Sequence of Events in the Learning Tasks (B) and in the Discrimination Task (C)



Note. (A) There were two phases: a Learning Phase and a Visual Discrimination Phase, administered in succession within a single experimental session. In the Learning Phase participants were trained to learn either four categories (category-learning group) or four pairings (paired-associate-learning group). The right-pointing arrows denote trained mappings of stimuli (on the left hand side) to responses (on the right hand side). Category learning involves multiple stimuli (with different sizes and border colors) mapping onto each response, whereas paired-associate learning involves a single (black-border) stimulus paired with each response. Two of the categories were denoted by verbal labels (label categories), whereas the other two were denoted by hard-to-name visual symbols (ideogram categories). Similarly, there were two verbal (label) pairings and two hard-to-name (ideogram) pairings. Following learning, there was a second (Visual Discrimination) phase, where all participants were administered an eyetracking visual discrimination task employing the previously trained shapes. Panels (B) and (C) depict the sequence of events during the trials of the Learning and Discrimination Sessions, respectively.

Labels have been suggested to selectively sensitize perceptual dimensions in a pervasive but transient fashion (Lupyan, 2021a; 2012b). That is, the ground upon which labels exert an influence is the sensitization of perceptual space. We reasoned that if the learning task does not involve such sensitization then no effect of labels should be observed. To this end, we employed paired-associate learning which—to our knowledge—has not been suggested to induce sensitization of perceptual dimensions (see General Discussion for more on this issue). Specifically, in our study another group of participants was administered a task that highly resembled the categorization task in every respect, but—

crucially—did not require the forming of categories; instead, it required the forming of pairs. In this paired-associate task there was only one perceptual dimension that varied across four stimuli (the dimension of shape, see Fig. 1A). The four stimuli had to be mapped to four labels, thus forming four pairs[1]. Similarly to the categorization task, the response cues for two of the pairs were pseudowords (label pairings) and the response cues for the other two pairs were visual symbols (ideogram pairings). If labels facilitate learning due to a perceptual-sensitizing mechanism (Lupyan, 2012a; 2012b) and paired-associate learning is not mediated by the same mechanism, then the label and the ideogram pairs should be learned with comparable accuracy. Alternatively, if labels facilitate learning in a more general way (Lupyan et al., 2007), then the label pairs should be learned with increased accuracy compared to the ideogram pairs.

Method

Participants

Sixty nine students of the University of Athens participated in exchange for course credit, meeting the requirements of normal or corrected-to-normal vision, no diagnosis of dyslexia, and Greek being their native language. All participants provided—verbally—informed consent. The experiment consisted of two phases (see Fig. 1). In the first phase participants were required to learn new categories (category-learning group) or new pairs (paired-associate-learning group). In the second phase both groups of participants were administered a visual-discrimination task involving eye tracking. The data from 21 participants were discarded prior to any data analysis due to difficulties with the eye-tracking procedure, e.g., reduced calibration accuracy caused by eye glasses or contact

[1] We are not implying that there is a minimum number of exemplars required for a category to be built. We are only arguing that a fixed shape that is repeatedly presented in the same "context" (here, the computer's white screen) is not a category.

lenses or technical failures. Thus, results reported here correspond to a sample of 48 students who completed both tasks. Twenty four (four male) of them were randomly assigned to the category-learning group (age *M* = 20.3 years; *SD* = 1.5) and 24 (three male) to the paired-associate-learning group (age *M* = 22.5 years, *SD* = 5.9).

Materials

Categorization Stimuli. Four four-point abstract shapes of low association value (and thus considered hard-to-name; Hulme et al., 2007; MacLeod & Dunbar, 1988) from the Vanderplas and Garvin (1959) repository were perceptually equated in size, using the method of adjustment (implemented in PsychoPy; Peirce, 2007) to obtain points of subjective equality (PSEs). Specifically, ten participants (not taking part in the main study) were asked to increase or decrease the size of the random shapes in order to match the size of a circle which remained constant in size. In each trial of this perceptual-equation task, two shapes were presented side-by-side on the screen, the circle and one of the random shapes. Each shape was presented ten times, half of them in an initial size that was greater than that of the constant shape. In each trial the screen side (left or right) and exact position of the stimuli were random. Points of subjective equality (PSEs) were calculated for each shape by averaging participants' average responses. Subsequently, 288 categorization items were created—72 for each shape—by varying the size (randomly within 0.2–0.8 of the PSE) and border color (randomly selected hues) of the PSEs.

Paired-Associate Stimuli. Four association items were created, identical to the abstract shapes used in category learning but with black margin and size corresponding to 75% of the PSEs. Categorization and paired-associate stimuli may be found at https://osf.io/rdnf7/.

Pseudowords. Two pseudowords served as response cues for the label categories and pairings, namely σάβης (/'savis/) and ρήτης (/'ritis/), previously used by Fotiadis and Protopapas (2014). The two pseudowords were equal in numbers of letters, syllables, and phonemes, stress position, and orthographic typicality (the mean orthographic Levenshtein distance of the 20 nearest neighbors—OLD20—was 2.00 for both pseudowords taking stress into account; Protopapas, Tzakosta, Chalamandaris, & Tsiakoulis, 2012; Yarkoni, Balota, & Yap, 2008).

Ideograms. Response cues for the ideogram categories and pairings were two hard-to-name Chinese characters (previously used by Fotiadis & Protopapas, 2014), namely 辛 (U+8F9B) and 辰 (U+8FB0). One stroke was erased from the second character to equate number of strokes—and thus perceptual complexity.

Procedure

Category Learning. Participants were administered 288 training trials, organized in 12 blocks of 24 trials. Each shape was presented equally often within a block. Participants never saw a categorization item twice.

Participants were told that they would be presented with four different shapes in varying size and color, and with four responses: two names and two ideograms. Their job was to learn which shape (disregarding color and size) corresponded to each response.

At each trial a fixation cross was presented for 500 ms, followed by a categorization item presented for 2000 ms. The two pseudowords and the two ideograms appeared next in a random vertical arrangement for a maximum of 10000 ms. Participants responded by clicking on a response option, and feedback—the words “correct” or “wrong” in Greek—was delivered for 500 ms. The procedure was programmed in DMDX display software (Forster &

Forster, 2003). Participants were given eight practice trials in the beginning and a short break after every four blocks. The task lasted approximately 35 minutes.

The order of categorization items and the permutation of response cues were pseudorandom (implemented with MIX; Van Casteren & Davis, 2006), with constrains precluding the same permutation of response cues in consecutive trials, and the same shape in more than two consecutive trials. All possible permutations of response cues were presented in each block. Participants received the same order of categorization items. Assignment of shapes to categories was counterbalanced across participants with the constraint that the shapes were paired, so that two shapes belonging to a pair were both predictive of either label or ideogram categories. This resulted in eight possible combinations of shape-response assignment, with three participants randomly assigned to each combination. To minimize participants' discomfort due to the head rest, eye-movements were not recorded during this session. DMDX scripts and details of the procedure (e.g., shape assignment), for both the category and the paired-associate learning task, can be found at https://osf.io/rdnf7/.

Paired-Associate Learning. The paired-associate learning task was a modification of the category-training task in that (a) each categorization item (varying in size and color) was replaced with the corresponding association item, and (b) instructions made no reference to either size or color.

Data Analysis

Analyses were conducted in R version 4.2.1 (R Core Team, 2021), employing generalized additive mixed models (Wood, 2011) with binomial distributions (Dixon, 2008), via a logit tranformation (Jaeger, 2008), fitted with restricted maximum likelihood and marginal likelihood estimation using package mgcv (Wood, 2011). Model comparison and

visualization of model estimates was done using package itsadug (Van Rij et al., 2020). Data and analysis scripts are available at https://osf.io/rdnf7/.

Results

No-response trials (three from the category-learning group and four from the paired-associate-learning group, comprising 0.05% of the data) and trials with response latencies less than 250 ms (five, from paired-associate-learning only; 0.03%) were excluded from analyses. Participants' accuracy increased, as trials progressed, in learning both the label and ideogram categories, averaging 88.8% (*SD* = 5.8) correct in category learning and 88.3% (*SD* = 5.6) in paired-associate learning (label categories: *M* = 90.5%, *SD* = 5.7; ideogram categories: *M* = 87.1%, *SD* = 6.8; label pairings: *M* = 87.9%, *SD* = 6.5; ideogram pairings: *M* = 88.4%, *SD* = 6.6). Figure 2 depicts participants' learning accuracy for the category-learning and the paired-associate-learning group in blocks of 24 trials.

Accuracy was analyzed using generalized additive mixed-effects models (Baayen et al., 2017). A null model with no fixed effects was first created, with a smooth term of trial as well as random effects modeling individual variability in the shape of the learning curve. In R notation the model was:

```
m0: acc ~ s(trial) + s(trial, sbj, bs = "fs", m=1)
```

with s(trial) denoting a smooth term of trial, and s(trial, sbj, bs = "fs", m=1) denoting by-participant random smooth terms of trial.

The null model (m0) was compared to a model (m1) that included the interaction of learning group (category vs. paired-associate) by condition (label vs. ideograms), and also kept the smooth term of trial, as well as random effects. In R notation, the model was:

```
m1: acc ~ group * condition + s(trial) + s(trial, sbj, bs = "fs", m=1)
```

Model comparison procedures indicated that m1 provided better fit to the data.

To test if learning progressed differently depending on group and/or condition, a dummy variable was created to encode the interaction of learning group by condition ("igc": Interaction of Group by Condition) in four distinct levels, namely category-labels, category-ideograms, paired-labels, and paired-ideograms. A model (m2) including four different smooth terms of trial, one for each level of the igc variable, was then fit:

```
m2: acc ~ group * cnd + s(trial) + s(trial, by = igc)
        + s(trial, sbj, bs = "fs", m = 1)
```

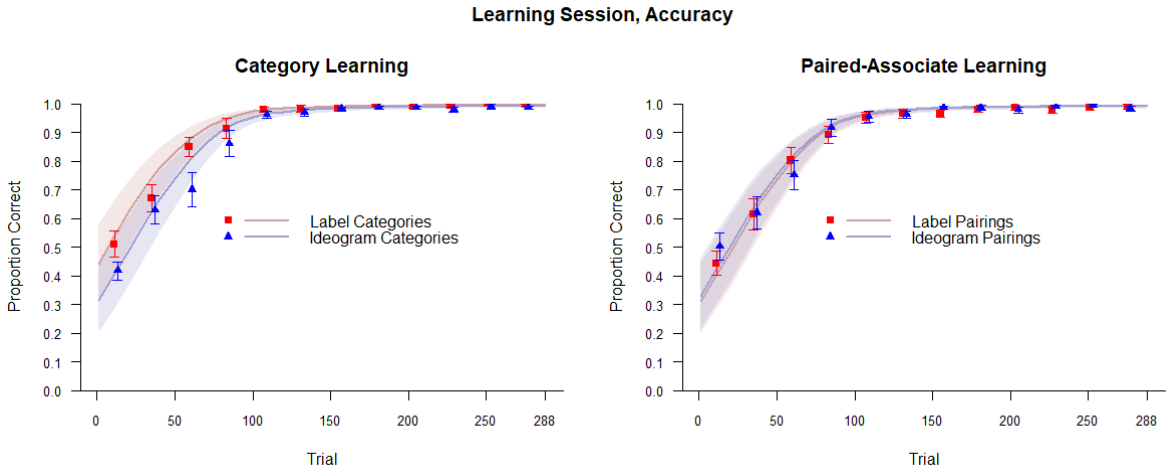
Model comparison procedures indicated that the term modeling different smooth terms (i.e., m2) did not improve model fit. Best-fit model estimates of learning accuracy are depicted in the line graphs in Fig. 2.

Our research question concerned the difference in accuracy between the label and ideogram categories, and also between the label and ideogram pairings. The fact that the best-fit model (m1) included a single smooth term of trial, independent of group and condition, means that any difference between learning label and ideogram categories, or label and ideogram pairings, remained constant throughout the training session. The estimate of the difference between label and ideogram categories suggested that labels provided facilitation: b = 0.548, *SE* = 0.097, z = 5.650, $p < .001$. In contrast, the estimate of the difference between the label and ideogram pairings (obtained by re-leveling the group factor and refitting the model) suggested that there was no facilitation in the paired associate learning group: b = −0.082, *SE* = 0.093, z = −0.890, p = 0.374. (These p values are not adjusted; the former survived Bonferroni correction for two comparisons.)

Additional analyses (available under Supplementary Analyses at https://osf.io/rdnf7) revealed that the two groups (irrespective of whether the response cues were labels or ideograms) did not differ in accuracy (b= −0.219, *SE* = 0.259, z = − 0.846, p = 0.397).

Figure 2

Results of the Learning Phase



Note. Points (squares and triangles) depict average accuracy of participants in learning the label and ideogram categories (left panel) and the label and ideogram pairings (right panel) in blocks of 24 trials. Error bars show between-subjects standard errors of the means. Smooth lines depict best-fit model estimates of accuracy, excluding random effects of participants. Error bands show 95% confidence intervals of the estimates.

Discussion

A label advantage was found during learning to categorize: participants exhibited increased accuracy in learning the label compared to the ideogram categories throughout the task (cf. Brojde et al., 2011; Tolins & Colunga, 2015). Moreover, this label advantage was found to be specific to the learning of categories: there was no facilitation due to labels during learning to associate. Overall, these results are supportive of the assumption of a labels-dependent dimension-sensitizing mechanism offering facilitation during learning (Lupyan 2012a; 2012b).

Study of Sustained Effects

Beyond the immediate effects observed during the learning process, labels were predicted to exert an influence after the categories have been learned. The mechanism of selective sensitization due to verbal labels has been suggested to account for phenomena of categorical grouping on the processing of categorization items of well-known categories. This influence is not an all-or-none phenomenon, but rather depends on the level of activation of participants' linguistic activity. Lupyan (2012a; 2012b) suggested that linguistic activity may be up-regulated (by, e.g., presenting labels at the beginning of each experimental trial, Lupyan & Thompson-Schill, 2012), allowed to exert an influence with no intervention (labels are self-activated when a categorization item is presented), or down-regulated (through, e.g., verbal interference, Lupyan, 2009; see Perry & Lupyan, 2013, for a critical review of such methodologies). By manipulating linguistic activity it has been shown that the effect of labels for well-practiced categories ("concepts") is a pervasive yet transient phenomenon in numerous test tasks that were utilized in this research program: visual search (Lupyan & Spivey, 2008), same–different discrimination (Lupyan, 2008b; Lupyan et al., 2010), picture verification (Edmiston & Lupyan, 2015; Lupyan & Thompson-

Schill, 2012; Perry & Lupyan, 2016), probe detection (Lupyan & Spivey, 2010b),"odd-one-out" procedures (Lupyan, 2009), object detection (Lupyan & Spivey, 2010a; Lupyan & Ward, 2013), and object recognition (Lupyan, 2008a). Thus, the effect of verbal labels for well known ("overlearned;" Lupyan & Spivey, 2010b) categories might be said to be well supported.

Sustained effects of labels should also be evident after newly-learned categories have been formed, since the label feedback hypothesis makes no distinction between new and overlearned categories. Tolins and Colunga (2015) tested this assumption by training participants to learn new artificial categories with or without the presence of redundant category labels. To test for long-lasting effects, the categorization stimuli were used in a subsequent categorization task. In this second task there was no label present, and the categorization rule changed. It was predicted that the category-diagnostic dimension should be sensitized to a greater extend when a label was present, and that this sensitization should be evident in the post-learning task because labels are self-activated. In a series of two experiments there was no evidence of label facilitation during learning, nor of the predicted sustained effects of labels in the post-learning task. Results showed that labels affected categorization only when the change in the categorization rule involved a reversal (i.e., stimuli that previously belonged to category A were assigned to category B, and vice versa), and this result was taken to suggest that verbal labels act as material symbols facilitating category-to-response mappings.

The fact that there was no evidence of sustained effects of labels for newly-learned categories constitutes a challenge for the label feedback hypothesis. Additionally, we argue, there was a limitation in the Tolins and Colunga (2015) study that needs to be addressed for a proper investigation of labels' sustained effects. In both of their experiments there was no

evidence of facilitation during learning[2], and, therefore, it seems reasonable to assume that labels for the categories did not selectively sensitize the perceptual space during learning in their experiments. It is therefore only natural to expect that self-activated labels would not warp perceptual space following learning, given that the same mechanism is proposed to underlie effects both during and following learning. Although it is unclear why labels had no effect in the Tolins and Colunga study, a possible explanation is that labels were not activated and thus did not exert an influence. For example, it could be argued that participants may have ignored the redundant labels in that particular experimental context. Consistent with this assumption, Lupyan (2006) provided preliminary evidence suggesting that labels do not provide an effect by just being present; they have to be learned by participants (see also Brojde & Colunga, 2011 for results supporting this argument). In our learning task participants could not ignore the labels since they were not redundant; they had to click on the category labels (either pseudowords or ideograms) in order to respond. In sum, given the importance of sustained effect for the viability of the label-feedback hypothesis, and also the limitations of the Tolins and Colunga study, further testing of such long-lasting effects of labels for newly-learned categories seems to be warranted.

We tested for sustained effects of labels by investigated the allocation of attention following learning. Perry and Lupyan (2016) suggested that selective attention may be thought of as the sensitization of perceptual space, influenced by verbal labels for the categories. We therefore predicted that shapes that had been predictive of named categories (label shapes) should capture attention to a greater extent compared to shapes

[2] In their Experiment 2 the result of no accuracy boosting during learning is to be expected, since the category-diagnostic dimension was orientation, which is not a typically diagnostic dimension. But in their Experiment 1 the category diagnostic dimension was frequency, and in this case redundant labels are predicted to boost accuracy during learning.

that had been predictive of hard-to-name categories (ideogram shapes). This effect should be evident in a post-learning task, because learned labels are naturally self-activated and therefore affect perceptual processes. Specifically, immediately after the learning session, both groups of participants performed a visual discrimination task on the trained shapes (in the absence of labels and ideograms, see Fig. 1B), while their eye movements were monitored (cf. Farrell, 1985; Belke & Meyer, 2002). Based on the findings of Rehder and Hoffman (2005a; 2005b), fixation durations were treated as measures of attention. In particular, we applied this rationale to the post-training discrimination task: Insofar as the shapes that had previously been predictive of named categories would capture attention to a greater extent, compared to the shapes that had previously been predictive of hard-to-name categories, participants should spend more time fixating the label shapes than the ideogram shapes in post-training trials presenting one label and one ideogram shape to be discriminated. In contrast, attention was assumed to be equally captured by shapes that had previously been associatively paired with either named or hard-to-name response cues, because associative learning is purportedly not mediated by the same sensitization mechanism as category learning. Therefore, participants in the paired-associate-learning group were predicted to spend comparable amounts of time fixating the label and ideogram shapes.

Method

Materials

The four association items used in the paired-associate learning task were presented as stimuli for the discrimination task. Each stimulus subtended a rectangle of roughly 2.5 cm horizontally by 6 cm vertically (1.8° × 4.4° of visual angle), presented on a 20-inch flat LCD monitor with a 1600 × 900 resolution at 60 Hz. Stimuli were placed 13.6 cm (10° of visual

angle) to the left and right of center, to minimize the effectiveness of peripheral vision (cf. Belke & Meyer, 2002) and encourage eye movements. Random jitter—both horizontally and vertically—of maximum ±20 pixels (0.4° of visual angle) for both stimuli introduced a slight uncertainty about exact position to prevent iconic-memory strategies from dominating performance and—again—to encourage eye movements.

Procedure

Participants were administered four blocks of 24 trials, programmed in Experiment Builder software (SR Research Ltd.). The script is available at https://osf.io/rdnf7/. Each stimulus was presented equally often within a block. Half of the trials were *different*, that is, presented two different stimuli on the screen, and the other half were *same*, that is, presented the same stimulus on both locations. All possible permutations of stimuli were included within a block. In pilot testing participants were found not to fixate a stimulus if it had just been presented on the same side of the screen. Thus, the pseudorandom trial order was constrained to preclude presentation of the same stimulus on consecutive trials (following Belke & Meyer, 2002).

A discrimination trial started with a drift check, followed by a fixation cross—subtending a square with a side of 1° of visual angle—presented for a minimum of 500 ms. Presentation of the two shapes was triggered by the participants' gaze recorded within the fixation cross for 150 ms, and lasted for 3000 ms. Participants were required to press one of two keys on the keyboard—as fast and accurately as possible—to denote whether the two shapes were different or the same.

An Eyelink 1000 Plus eyetracker sampling monocularly at 2000 Hz recorded the eye providing the best calibration accuracy. A head and chin rest was used, and calibration took place on average every two blocks, or more often if required. Participants were given a

block of practice trials, there was a short break in the middle of the procedure, and the task lasted on average 20 minutes.

Data preprocessing and analysis

Analyses of participants' behavioral measures, including accuracy and response latencies, were exploratory, since no predictions were made concerning these measures. Accuracy was analyzed by analysis of variance. Response latencies were best fit by the gamma distributions and were analyzed with generalized mixed models.

Analyses of participants' eye movements focused on fixation duration and only included data from trials presenting one label and one ideogram shape (8 trials within each block). The online parser of SR Research Ltd was used for fixation detection. Two rectangular areas of interest (AOIs) were defined prior to data collection, each subtending a square exceeding each stimulus by a margin of 5.5° of visual angle (7.5 cm). This margin was defined as the sum of the equipment's nominal accuracy (0.5° of visual angle) and a rough measure of the span of peripheral vision (5° of visual angle). Because of the substantial eccentricity of stimulus placement near the screen edges, the AOIs were not symmetric (only 100 pixels, amounting to 2.78 cm, or 2° of visual angle, for the distal margins). Duration of fixations within an AOI was determined using the Data Viewer software (SR Research Ltd.).

Following Henderson et al. (1999; Võ & Henderson, 2009), fixations with duration less than 90 ms or greater than 1000 ms were excluded from analysis. The sum of durations of fixations landing within an AOI (hereafter "dwell time") was calculated for each participant and trial. Dwell time was analyzed as a function of trial, to account for temporal order effects (such as increasing familiarity; cf. Lupyan & Spivey, 2008, Supplementary Materials), using generalized linear mixed-effects models with gamma distributions.

Confidence intervals for the model's fixed effects estimates were extracted using package merTools (Knowles & Frederick, 2020). To more thoroughly investigate the time course of the predicted difference between label and ideogram shapes, dwell time was additionally analyzed with generalized additive mixed models (Wood, 2011) with gamma distributions, fitted with restricted maximum likelihood and marginal likelihood estimation using package mgcv (Wood, 2011). Model comparison and visualization of model estimates was done using package itsadug (Van Rij et al., 2020). All analyses were conducted in R version 4.2.1 (R Core Team, 2021), and data along with analysis scripts may be found at https://osf.io/rdnf7/.

Results

Accuracy

No-response trials (11 from the category-learning group and 5 from the paired-associate-learning group, totaling 0.34% of the data) were excluded from analysis. Both groups were highly accurate in the task: Taking all trials into account, participants averaged 98.1 % (*SD* = 1.9) correct responses in the category-learning group and 98.3 % (*SD* = 1.6) in the paired-associate group. A Welch's t-test for independent samples suggested that the two groups did not differ in accuracy, $t(44.08) = -.28$, $p = .783$. To examine the differences between *same* and *different* trials we conducted a two-way ANOVA on average accuracy, with group as the between-subjects factor, and type of trial (*same* vs. *different*) as the within-subjects factor. Results revealed no interaction of group by type, $F(1, 46) = 1.392$, $\eta^2 = .014$, $p = .244$, and no effect of group, $F(1, 46) = .075$, $\eta^2 = .0008$, $p = .785$. There was an effect of trial type, $F(1, 46) = 6.769$, $\eta^2 = .066$, $p = .013$, suggesting higher accuracy in *different* than in *same* trials, consistent with previous findings (Farell, 1985). Further analyses (available at https://osf.io/rdnf7/) revealed no effect of response cues on accuracy.

Response Latencies

No-response trials (11 from the category-learning group and 5 from the paired-associate-learning group, totaling 0.34% of the data) as well as incorrect trials (43 from the category-learning group and 40 from the paired-associate-learning group, totaling 1.80% of the data) were excluded from analysis. Mean response latencies were 1081.71 ms (*SD* = 210.67) in the category learning group and 962.92 ms (*SD* = 148.70) in the paired-associate group. This difference was significant, $t(41.361) = 2.263$, $p = 0.029$. Within each learning group, participants responded equally fast to same and different trials (ps > .05).

Analysis of response latencies was done using generalized mixed-effects models with gamma distributions. Model comparison procedures suggested that the best-fit model was:

```
m_rt: RT ~ trial*Group + trial*same_diff + (1|sbj)
```

with (1 | sbj) denoting by-participant random intercepts. There was a significant interaction of trial by group: $b = 0.584$, $SE = 0.244$, $t = 2.40$, $p = .017$. Visualization of the model's estimates suggested that participants in the category learning group took longer to respond compared to the paired-associate group at (both for same and different trials), and that both groups' speed of responding increased as trials progressed. There was also an interaction of trial by type of trial (same vs. different): $b = -1.208$, $SE = 0.156$, $t = -7.773$, $p < .001$. Visualization suggested that participants took longer to respond to same than to different trials at the beginning of the task. Participants' speed of responding increased for both types of trials and eventually converged towards the end of the task. All analyses and visualizations are available at https://osf.io/rdnf7/.

Average Dwell Time

This analysis included only critical trials, that is, *different* trials displaying one label shape and one ideogram shape (eight per block). Participants in the category-learning group

spent on average 668.28 ms (*SD* = 213.09) fixating the label shapes and 683.82 ms (*SD* = 220.49) fixating the ideogram shapes. Participants in the paired-associate-learning group spent on average 726.59 ms (*SD* = 225.21) fixating the labels shapes and 727.52 ms (*SD* = 205.96) fixating the ideogram shapes. Participants in both groups spent on average comparable time fixating the shapes (Category learning: 676.04, SD = 210.57, Paired-associate: 726.95, SD = 201.437) as indexed by a t-test ($t(45.91) = -0.856$, $p = .397$).

Analyses suggested that the gamma distribution fit the data better than the normal distribution. In R notation, the model used was:

```
m: dwell_time ~ group * cnd * trial + (1+cnd*trial|sbj)
```

with (1+cnd*trial|sbj) denoting by-participant random intercepts and by-participant random slopes of trial interacting with condition (label vs. ideogram shape). Trial was centered to facilitate convergence. Results showed a triple interaction of group (Category vs. Paired-Associate learning) by condition (label vs. ideogram shape) by trial, $b = 32.545$, $SE = 7.055$, $t = 4.613$, $p < .001$.

To pinpoint the locus of the significant difference, we next analyzed data from the category-learning group only:

```
m_cat: dwell_time ~ cnd * trial + (1+cnd*trial|sbj)
```

Results showed an interaction of condition (label vs. ideogram shape) by trial for the category-learning group: $b = -47.26$, $SE = 13.29$, $t = -3.555$, $p < .001$, and no fixed effects of either condition or trial (ps > .05).

A similar model was fit to data from the paired-associate-learning group only:

```
m_pa <- dwell_time ~ cnd * trial + (0+cnd*trial|sbj)
```

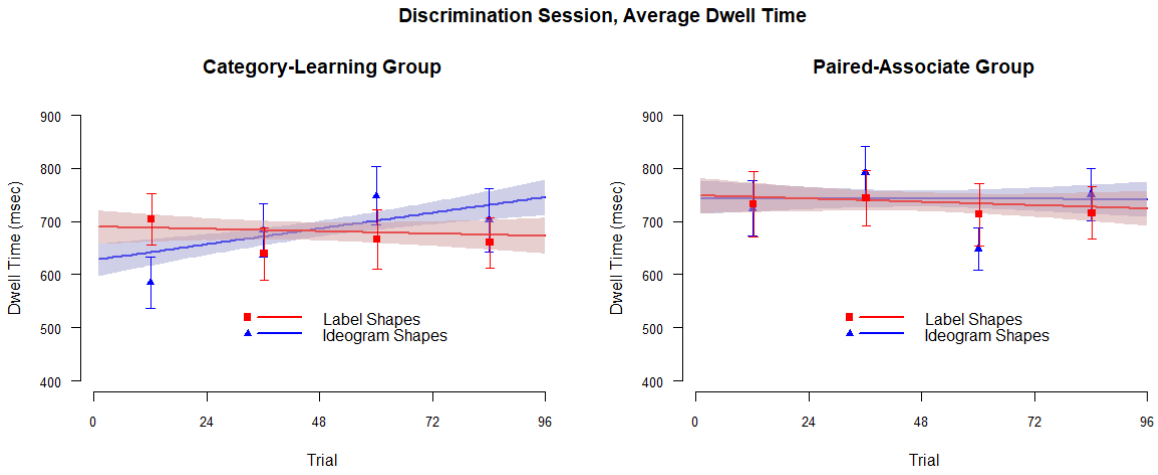
Full random structure caused convergence issues, therefore in this model only by-participant random slopes of trial interacting by condition were included. In this model

there was no interaction of trial by condition: b = 3.048, SE = 19.90, t = 0.153, p = .878, and no main fixed effects (ps > .05).

Estimates of the full model (m), excluding random effects groups of participants, are shown in Fig. 3, separately for the two groups of participants. As previously noted, to further investigate if there was a difference in dwell time between the label and ideogram shapes, we analyzed data from each group using generalized additive models. These models indicate the range of values of an independent variable—in our case trial—for which a difference is significant. There was no difference during the task in dwell time between shapes that had previously been predictive of named or hard-to-name categories. Similarly, there was no difference in dwell time between shapes that had previously been paired with verbal labels or hard-to-name response cues. All analyses are available at: https://osf.io/rdnf7/.

Figure 3

Results of the Discrimination Session



Note. Points (squares and triangles) depict average dwell time in the label and ideogram shapes for the category-learning group (left panel) and the paired-associate-learning group (right panel) in blocks of eight trials (eight out of 24 trials in each block were the trials of interest). Error bars show between-subjects standard errors of the means. Lines depict the model's estimate of dwell time, excluding random effects of participants. Error bands show 95% confidence intervals of the estimates.

Discussion

In the Discrimination Phase of our study, we examined the sustained effects of category labels on attention, contrasted with the sustained effects of labels for associations. Results suggested that the deployment of attention (indexed by fixation durations) as trials progressed was dependent on whether the shapes had previously been predictive of named or hard-to-name categories. In contrast, attention was similarly captured by shapes that had previously been paired with either labels or hard-to-name symbols in paired-associate

training. Our results do not clearly bear out our prediction of increased fixation durations on the label shapes compared to ideogram shapes for the category-learning group. Nevertheless, and in contrast to previous studies (Tolins & Colunga, 2015), the present findings are indicative of sustained effects of labels on attention mechanisms, importantly, only after learning to categorize and not after learning to associate. The finding of an interaction of trial progression by type of shape on the deployment of attention for the category learning group suggests that the sustained effects of labels may be subject to participants' adaptation to the task, an issue considered further in the General Discussion.

General Discussion

Theories of the interaction between linguistic and perceptual processes suggest that verbal labels for the categories selectively sensitize perceptual dimension that are typically diagnostic in everyday categorization (Lupyan 2012a; 2012b; Perry & Lupyan, 2014). This mechanism is predicted to provide both immediate facilitative effects during learning and is also supposed to affect attention processes following learning (Lupyan et al., 2007; Tolins & Colunga, 2015). Previous research, though, provided mixed results (Brojde et al, 2011; Lupyan & Casasanto, 2015; Tolins & Colunga, 2015), necessitating further exploration of the immediate and sustained effects of labels.

In our study a group of participants learned named and hard-to-name artificial shape-based categories. To preclude alternative explanations of a purported labels effect, we recruited an additional group of participants who were asked to learn named and hard-to-name shape-based pairings. To examine the sustained effect of labels on attention, both groups of participants were subsequently administered a visual discrimination task while their eye movements were monitored.

Comparing Category and Paired-Associate Learning

A comparison of learning effects for categories with learning effects for associations is not novel. A similar approach was used by Poldrack et al. (2001) in a neuroimaging study aiming to shed light on brain structures employed specifically during the learning of categories (and not learning in general). Nevertheless, it is not a common approach in the field, thus warranting further scrutiny. In the present section we clarify the rationale of this comparison and its relevance to our research questions.

Our experimental design is based upon the claim that the ground upon which verbal labels exert an influence is the warping of perceptual space (see Lupyan, 2012b). We assume that when participants perform the paired-associate task in our study perceptual space is not warped in the same way. Therefore, comparing the effect of labelling for categorization and labelling for associations is of interest. To support this argument, we aim to rule out two possibilities: First, that perceptual warping also occurs during and following the paired-associate task due to a categorization-like mechanism. Second, that learning named pairings might induce perceptual warping due to alternative mechanisms.

Perceptual warping occurs when categorizing *or* when naming an entity, because naming is, in effect, categorization (Goldstone, 1994; Lupyan, 2012b). Indeed, names can serve as category labels (Goldstone et al., 2001); when naming an entity or an object, various perceptual aspects, as experienced over different occasions in the course of time (such as orientation, luminance, background stimuli, or even facial expressions if naming an animate entity), should be ignored/desensitized whereas other features are emphasized. So, the effects of labels are observed following the learning of a name through a categorization-like process arguably affecting perceptual space (Lupyan et al., 2020). One can argue that learning named pairings in our paired-associate task is similar to learning to name an object

and could be accompanied by the warping of perceptual space. However, we contend that our paired-associate regime is not equivalent to learning to name an object and bears no resemblance—in terms of cognitive demands—to naming an object as it happens in everyday life. Specifically, our participants were only presented with a unique stimulus constant in all perceptual dimensions. There was thus no "irrelevant" perceptual dimension to be ignored or deemphasized and no perceptual warping to be induced though a categorization-like mechanism based on abstracting over and emphasizing perceptual dimensions (Goldstone, 1994).

In considering alternative possibilities, besides categorization-like mechanisms, we are not aware of any evidence that learning to pair shapes to names—in an one-to-one fashion—involves perceptual warping of the shape dimension. Admittedly, this issue has not been thoroughly investigated. A recent study is perhaps most relevant: Calignano et al. (2021) employed a training regime similar to our paired-associate training task. They showed that pairing a novel object with an auditory pseudoword induced *reduced* capturing of attention compared to an object-only condition or compared to pairing with non-linguistic sounds. This finding may be taken to suggest that, during paired-associate learning, labels affect attention in the *opposite direction* compared to category learning, thus supporting the argument that our paired-associate task constitutes a valid approach to testing the label-feedback hypothesis.[3] In sum, the literature has only provided preliminary results but no indication that paired-associate learning, as implemented in our experiment, should induce perceptual warping the way category learning is predicted to do. We thus

[3] Caution is warranted in interpreting this finding, however, because presenting the object with the previously paired pseudoword had a similar effect on attention as presenting it with an irrelevant pseudoword, leaving open the possibility that attention was affected by the presence of any linguistic sound, irrespective of what was previously learned.

conclude that examining the effects of labels when learning to associate constitutes a theoretically relevant contrast to category learning. The validity of this contrast is also supported (post-hoc) by our results in showing different effects of labels depending on training regime.

Immediate Effects of Labels

To examine the effects of labels on learning to categorize, we manipulated the nameability of the response cues. Contrary to previous studies (Brodje et al., 2011; Tolins & Colunga, 2015) our experiment revealed a label advantage during learning to categorize, in that named categories were learned with increased accuracy compared to hard-to-name categories. Importantly, this advantage cannot be attributed to special selection of experimental material (c.f. Casasanto & Lupyan, 2015) given our counterbalanced materials and procedures. Moreover, this effect cannot be attributed to a general facilitation due to the processing of verbal stimuli (Lupyan et al., 2007) since there was no effect of labels during learning to associate. Overall, these results showed a label advantage during learning to categorize (Lupyan et al., 2007) and are supportive of a labels-dependent mechanism inducing perceptual sensitization of category-diagnostic dimensions (Lupyan, 2012a; 2012b).

Sustained Effects of Labels

Following learning to map stimuli to named or hard-to-name response cues, participants in our study were given a visual discrimination task, while their eye movements were monitored. Participants fixated shapes that were diagnostic of named categories differently compared to shapes that were diagnostic of hard-to-name categories, as indicated by an interaction of trial progression by type of shape on dwell time. Importantly, participants in the paired-associate group fixated shapes that were previously paired to either named or hard-to-name symbols similarly throughout the task. Although this pattern

was not predicted, these results are of interest as they constitute the first piece of evidence suggesting sustained effects of labels for newly-trained categories. Importantly, these sustained effects were observed along with facilitative immediate effects of labels during learning, consistent with the idea that both immediate and long-term effects of labels are the product of the same mechanism inducing the warping of perceptual space.

The finding of a dynamic deployment of attention—dependent on verbal labels—as trials progressed, is consistent with experience from previous explorations of the idea that visual processing—affected by labels of overlearned categories—might depend on experimental trial (Lupyan & Spivey, 2008). It is also consistent with the finding that the effect of overtly presenting labels of categories is time-dependent both for overlearned (Lupyan & Spivey, 2010b) and newly-familiarized categories (Barnhart et al., 2018). Moreover, analysis of response latencies in the discrimination task revealed that participants' speed of responding increased as trials progressed. Given the tight coupling of behavioral and eye-movement measures (Rehder & Hoffman, 2005a; 2005b), it seems plausible to also expect practice effects in fixation durations.

Local Sensitization of Dimensions

Our results suggest that, during learning, shapes that are diagnostic of named categories are sensitized to a greater extent compared to shapes that are diagnostic of hard-to-name categories. Following learning, shapes that were previously diagnostic of named or hard-to-name categories were found to capture attention differently as trials progressed. We argue that these results are important, also because they are only possible if a perceptual dimension may be selectively sensitized. That is, the results speak to the issue of whether labels selectively activate *specific values* of a dimension rather than the entire dimension.

This issue is far from trivial. The majority of experimental research examining category-learning processes and systems has utilized between-subjects manipulations and corresponding comparisons (Ashby & Maddox, 2005). Although this approach has proven fruitful in advancing our understanding of category learning, it does not help elucidate whether it is entire perceptual dimensions or, rather, specific perceptual values that are important for categorization. A perceptual dimension encompasses all possible values within it; therefore, even if it is the features of specific values that capture attention during learning to categorize, a between-subjects design is—in principle—not diagnostic of the distinction and can only attest in favor of dimensional sensitization or activation.

Surprisingly few studies have addressed the dimension vs. values distinction. As noted, Goldstone (1994) showed that perceptual sensitization following learning to categorize is a localized phenomenon (i.e., it is greater for values of a diagnostic dimension that cross a category boundary compared to values that belong to the same category). This result was replicated by Van Gulick and Gauthier (2014). In related research, Aha and Goldstone (1992) provided evidence suggesting that, following learning to categorize, different values of a perceptual dimension may be selectively attended to (see also Blair et al., 2009).

With respect to the label-feedback hypothesis, Lupyan (2012b) explicitly posited that it is specific perceptual features that are selectively activated by verbal labels, rather than general perceptual dimensions. However, the studies examining the initial and sustained effects of category labels (Brodje et al., 2011; Lupyan & Casasanto, 2015; Lupyan et al., 2007; Perry & Lupyan, 2014; Tolins & Colunga, 2015) have all used between-subjects manipulations. In contrast, in our experiments the varying nameabilty of formed categories was a within-subjects manipulation. All participants learned both named and hard-to-name

categories in a single training procedure. If labels activate the diagnostic perceptual dimension as a whole, rather than the features of specific diagnostic values linked to labels (i.e., the dimension of shape rather than the label shapes specifically), then we should have observed no difference in accuracy between label and ideogram category learning, as well as no difference in the post-learning processing of label and ideogram shapes. We may therefore conclude that labels for the categories result in increased activation (both during and also following learning) of specific values within a dimension, rather than activating the whole dimension, in accordance with Goldstone's (1994) suggestion that perceptual space is *locally* warped as a result of learning to categorize.

Interpretation under Alternative Theories of Category Learning

Our findings of the differential effects of easy-to-name vs. hard-to-name category labels have so far been interpreted exclusively through the label-feedback hypothesis framework, which was the theoretical foundation of our design and our predictions. Here we consider the implications of our results in light of other theoretical perspectives.

The warping of perceptual space as a result of learning to categorize is well documented (e.g., Goldstone, 1994; Goldstone et al., 2001; Pothos & Reppa, 2014). In addition, the development of functional features that reflect representational changes (and not just changes at a higher, decisional level; Schyns & Rodet, 1997) has been hypothesized to mediate learning effects and explain differences between novice and expert categorizers (Schyns et al., 1998). Under this framework, our participants may have learned to rely on parts of the stimuli (e.g., sharp edges on two shapes or almost right angles in the other two) to distinguish four shape-based categories. Future research should investigate the extent to which easy-to-name vs. hard-to-name labels may modulate the development of functional features.

The COVIS model of category learning (Ashby & Maddox, 2005) suggests that learning may be mediated by either the explicit system (employing hypothesis-testing processes) or the implicit system (employing information-integration processes). Both systems underlie learning simultaneously during a task and one of the two systems provides the response depending on trial demands (e.g., Ashby & Crossley, 2010). In our experiment, learning hard-to-name categories may have been mediated by the implicit system, resulting in lower accuracy, whereas easy-to-name categories were learned through the explicit system (a verbal rule being more plausible), resulting in higher accuracy (see Fotiadis & Protopapas, 2014, for an examination of this hypothesis based on the nameability of the categorization items). The COVIS framework does not involve representational change (Ashby et al., 1998), however, making it unclear how it might account for sustained effects of labels on attention during a post-categorization task.

Goldstone et al. (2001) examined the effects of learning to categorize on item similarity judgements. They provided evidence not only for representational change but also a strategic "label-based bias" (p. 30) in that items that share a category label are likely to be judged as being more similar to each other compared to items from different categories. Goldstone et al. did not suggest that these effects influence the learning process by modulating accuracy. However, in our study, hard-to-name labels may arguably have exerted smaller strategic influences compared to easy-to-name labels, differentially affecting category learning (which is driven largely by similarity, e.g., Nosofsky, 1986) without requiring an assumption of sensitization. It remains to be investigated if such strategic influences may manifest themselves in subsequent viewing behavior, and more specifically if they can account for the interaction of dwell time by category nameability.

The distinction between supervised and unsupervised categorization (Love, 2002; Pothos et al., 2011) may also be relevant to our findings. Pothos et al. suggested that the processes that support supervised categorization are intimately related to those of unsupervised category formation, both being based on similarity. Yet only supervised categorization involves "transformation of representations" (p.1710). Under this framework, hard-to-name category labels might have led to less supervised learning, and thereby less pronounced warping of psychological space, compared to easy-to-name categories. This account leads to predictions that are compatible with those of the label-feedback hypothesis (Lupyan, 2012a), both for initial and sustained effects, so further research is required to disentangle the two possibilities.

Finally, Pothos and Reppa (2014) examined the factors that modulate sensitization (operationalized as "similarity change") and found that similarity change was more pronounced for more difficult/less intuitive category structures compared to easier categories. This finding is relevant for our study in that the label categories were arguably more intuitive/easy than the ideogram categories, therefore (a) more accurately learned, and (b) leading to less pronounced sensitization involving the label shapes than the ideogram shapes. In contrast, Lupyan's (2012a; 2012b) prediction would be that more nameable labels entail greater sensitization, resulting in greater accuracy for easy-to-name categories than for hard-to-name categories. That is, both accounts suggest that easy-to-name categories are learned more accurately than hard-to-name categories, so accuracy cannot discriminate between the two. Predictions diverge regarding sustained effects, however, due to the contrasting hypotheses for sensitization: Pothos and Reppa's account would predict (under the assumption that greater sensitization results in increased attention; Perry & Lupyan, 2016) that label shapes should be less attended to than ideogram

shapes. In contrast, the prediction from Lupyan's account would be that label shapes should capture attention to a greater extent than ideogram shapes. Our finding of an interaction between type of category (easy-to-name vs. hard-to-name) and trial progression on dwell time during the post-categorization task cannot distinguish between the two accounts. Perhaps a training regime specifically designed to equate accuracy between easy-to-name and hard-to-name categories might be more informative, aiming to examine if attentional capture is promoted or hindered for items that had previously been diagnostic of hard- vs. easy-to-name categories.

In sum, we suggest that our paradigm of simultaneously learning easy-to-name and hard-to-name categories may provide testing benchmarks for theories of category learning. Comparing the effects of verbal vs. non-verbal labels for the categories may elucidate the nature of changes taking place at the representational level of the cognitive system.

Limitations and Future Directions

In examining the label-feedback hypothesis, we manipulated linguistic activity by using names vs. hard-to-name symbols, rather than by using redundant labels vs. the absence of labels (e.g., Brojde et al., 2011; Lupyan et al., 2007; Tolins & Colunga, 2015). This manipulation took the effect of correlated cues out of the equation but introduced a possible limitation. Verbal labels and ideograms were equated in size but arguably placed different demands on, e.g., memory or perception, potentially leaving the results open to alternative interpretations. Similar asymmetries are seen in previous studies (for example, between geometric and resistant-to-verbalization stimulus features, Kurtz et al., 2013, or between verbal labels and location cues, Lupyan et al., 2007), as it is not always clear what should be equated and by which criteria. Further theoretical and experimental work should address criteria and procedures for equating verbal and hard-to-name stimuli.

In our study we found initial effects and some indication of sustained effects of labels for the newly-learned categories on the processing of categorization items (cf. Tolins and Collunga, 2015, examining sustained effects in the absence of initial effects). The label-feedback hypothesis posits that the same mechanism underlies both effects during and also following learning (Lupyan, 2012a; 2012b) and it has been argued that the mere presence of labels does not suffice, but rather the labels have to be learned (Lupyan, 2006). We therefore submit that further investigation of the effects of redundant labels should include an assessment of the degree to which participants have learned the labels prior to examination of label effects, either initial or sustained. Alternatively, we suggest that using a paradigm like the one introduced here, which uses named and hard-to-name response cues, might be an efficient way of examining the effect of labels. Participants may not ignore the labels in such a procedure, and labels are, therefore, allowed to exert their influence on perceptual space.

In our study we contrasted the effects of labels for categories with those for associations. The two tasks differed only on which stimulus dimensions participants were exposed to. Categorization items varied in border color, size, and shape, whereas stimuli in the paired-associate task only varied in shape. In both tasks shape was the diagnostic dimension. We hypothesized that, although strikingly similar, the two tasks would affect perceptual space differently. In particular, perceptual space was predicted to be warped for category learners, but not for learners of associations. We have offered no direct evidence sensitization occurred during learning to categorize but not during learning to associate. Post-hoc multidimensional scaling analyses of response times from the discrimination task (available under Supplementary Analyses in https://osf.io/rdnf7/) are consistent with the hypothesis that the perceptual space following categorization is sensitized depending on the

nameability of labels, compared to learning to associate. Moreover, our findings of initial and some sustained effects for categories—but not for associations—seem to corroborate our hypothesis. Nevertheless, no firm conclusions can be drawn, at least from our experiment, given that we did not have a pre-learning vs. post-learning design to directly test for sensitization. Future studies should better investigate the nature of differences between the two tasks.

As noted, following learning, there was some indication of sustained effects of labels for the categories, but not for the associations, on the visual processing of learned shapes as indexed by fixation durations. An unanticipated result also emerged whereby participants who underwent category training were slower, during the post-training task, in deciding whether the two shapes were same or different, compared to participants who underwent paired-associate training. We submit that this result—though not specifically predicted—indicates that processes mediating category and paired-associate learning may differently affect post-learning behavioral measures. As an additional theoretical possibility, this result might reflect exemplar novelty during the test task: Participants in the paired-associate group had already seen the exact same stimuli of the discrimination task, whereas participants in the category-learning group had seen different exemplars (in size and color). However, we believe this possibility to be remote because the sizes of the test task stimuli were well within the established category of the category-learning participants (based on the space spanned by the exemplars). Still, this minor discrepancy between the two groups was the reason our predictions concerned sensitization effects within groups (label vs. ideogram shapes for each of the groups) and not strict between-groups comparisons. More research is required to further dissect the result of greater response latencies for the paired-associate group and disentangle the effects of novelty from the effects of learning regime.

Perhaps a task designed to track pupil size[4] (known to be indicative of attention processes; Kahneman, 1973) while participants view the trained shapes (equated for novelty) following learning could help shed light on the mechanisms allowing labels to influence the deployment of attention.

Conclusions

In the present study we investigated the effect of labels for the categories during initial learning and in a post-learning test task. We found that participants were more accurate in learning named compared to hard-to-name categories. It was also revealed that there was no label advantage during learning to associate, precluding explanation by a general theory of facilitation in processing verbal stimuli. Contrary to previous research (Tolins & Colunga, 2015), there was some evidence of sustained effects of category labels on attention mechanisms recruited in an eye-tracking visual discrimination task. These sustained effects did not emerge following learning to associate, attesting to an explanation based on a labels-dependent mechanism selectively sensitizing perceptual dimensions.

The present research contributes to the category-learning literature by suggesting that linguistic representations interact with perceptual and attention processes recruited both during and also following the learning of categories (Lupyan, 2012a; 2012b). Our results have challenging implications for current theories of learning, and also for the language and thought debate (e.g., Gleitman & Papafragou, 2013; Lupyan et al., 2020; Regier et al., 2010), helping elucidate the more general question regarding the interplay between the language faculty and learning processes.

[4] Although pupil size data were collected during the discrimination task, the design of the task, mainly the eccentricity of the stimuli and the variation in their placement, prevents a meaningful analysis (Mathôt & Vilotijević, 2022).

Acknowledgements

We thank Laura Ziaka for multi-faceted help in carrying out this research, Petros Roussos for help with recruiting participants, Sam Hutton for technical assistance with eyetracking, Victor Kuperman for statistical advice, Raymond Bertram for advice on preprocessing eyetracking data, and Simon Wood, Torsten Hothorn, Kenneth J. Kurtz, Lynn Perry, Jackson Tolins, Melissa Võ, and Antje Nuthman for responding to queries. Part of this work has been presented at the 4th Panhellenic Conference in Cognitive Psychology, Athens, Greece. Material, data, and analysis scripts are available at https://osf.io/rdnf7/.

Declaration of Conflicting Interests

The Authors declare that there is no conflict of interest

References

Aha, D. W., & Goldstone, R. L. (1992). Concept learning and flexible weighting. In *Proceedings of the 14th Annual Conference of the Cognitive Science Society* (pp. 534–539). Hillsdale, NJ: Erlbaum

Ashby, F. G., Alfonso-Reese, L. A., Turken, A. U., & Waldron, E. M. (1998). A neuropsychological theory of multiple systems in category learning. *Psychological Review, 105*(3), 442–481. https://doi.org/10.1037/0033-295X.105.3.442

Ashby, F. G., & Maddox, W. T. (2005). Human Category Learning. *Annual Review of Psychology, 56*, 149–178. doi:10.1146/annurev.psych.56.091103.070217

Belke, E., & Meyer, A. S. (2002). Tracking the time course of multidimensional stimulus discrimination: Analyses of viewing patterns and processing times during "same"–"different" decisions. *European Journal of Cognitive Psychology, 14*(2), 237–266. https://doi.org/10.1080/09541440143000050

Barnhart, W. R., Rivera, S., & Robinson, C. W. (2018). Effects of linguistic labels on visual attention in children and young adults. *Frontiers in psychology, 9*, 358. https://doi.org/10.3389/fpsyg.2018.00358

Baayen, H., Vasishth, S., Kliegl, R., & Bates, D. (2017). The cave of shadows: Addressing the human factor with generalized additive mixed models. *Journal of Memory and Language, 94*, 206-234. https://doi.org/10.1016/j.jml.2016.11.006

Blair, M. R., Watson, M. R., Walshe, R. C., & Maj, F. (2009). Extremely selective attention: Eye-tracking studies of the dynamic allocation of attention to stimulus features in categorization. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 35*(5), 1196–1206. https://doi.org/10.1037/a0016272

Boutonnet, B., & Lupyan, G. (2015). Words jump-start vision: A label advantage in object recognition. *The Journal of Neuroscience, 35*(25), 9329–9335. https://doi.org/10.1523/JNEUROSCI.5111-14.2015

Brojde, C. L., Porter, C., & Colunga, E. (2011). Words can slow down category learning. *Psychonomic Bulletin & Review, 18*(4), 798–804, https://doi.org/10.3758/s13423-011-0103-z

Calignano, G., Valenza, E., Vespignani, F., Russo, S., & Sulpizio, S. (2021). The unique role of novel linguistic labels on the disengagement of visual attention. *Quarterly Journal of Experimental Psychology,* Advance online publication. https://doi.org/10.1177/17470218211014147.

Dixon, P. (2008). Models of accuracy in repeated-measures designs. *Journal of Memory and Language, 59*(4), 447–456. https://doi.org/10.1016/j.jml.2007.11.004

Edmiston, P., & Lupyan, G. (2015). What makes words special? Words as unmotivated cues. *Cognition, 143*, 93–100. https://doi.org/10.1016/j.cognition.2015.06.008

Farell, B. (1985). “Same”–“different” judgments: A review of current controversies in perceptual comparisons. *Psychological Bulletin, 98*(3), 419–456. https://doi.org/10.1037/0033-2909.98.3.419

Folstein, J. R., Gauthier, I., & Palmeri, T. J. (2012). How category learning affects object representations: Not all morphspaces stretch alike. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 38*(4), 807–820. https://doi.org/10.1037/a0025836

Folstein, J. R., Palmeri, T. J., & Gauthier, I. (2013). Category learning increases discriminability of relevant object dimensions in visual cortex. *Cerebral Cortex, 23*(4), 814–823. https://doi.org/10.1093/cercor/bhs067

Folstein, J. R., Palmeri, T. J., & Gauthier, I. (2014). Perceptual advantage for category-relevant perceptual dimensions: the case of shape and motion. *Frontiers in Psychology, 5*, 1394. https://doi.org/10.3389/fpsyg.2014.01394

Folstein, J. R., Palmeri, T. J., Van Gulick, A. E., & Gauthier, I. (2015). Category learning stretches neural representations in visual cortex. *Current Directions in Psychological Science, 24*(1), 17–23. https://doi.org/10.1177/0963721414550707

Forster, K. I., & Forster, J. C. (2003). DMDX: A Windows display program with millisecond accuracy. *Behavior Research Methods, Instruments, & Computers, 35*(1), 116–124. https://doi.org/10.3758/BF03195503

Fotiadis, F. A., & Protopapas, A. (2014). The effect of newly trained verbal and nonverbal labels for the cues in probabilistic category learning. *Memory & Cognition, 42*(1), 112–125. https://doi.org/10.3758/s13421-013-0350-5

Fulkerson, A. L., & Waxman, S. R. (2007). Words (but not tones) facilitate object categorization: Evidence from 6-and 12-month-olds. *Cognition, 105*(1), 218–228. https://doi.org/10.1016/j.cognition.2006.09.005

Gleitman, L. R. & Papafragou, A. (2013). Relations between language and thought. In D. Reisberg (Ed.), *Handbook of Cognitive Psychology* (pp. 504–523). New York, NY: Oxford University Press.

Goldstone, R. L. (1994). Influences of categorization on perceptual discrimination. *Journal of Experimental Psychology: General, 123*(2), 178–200. https://doi.org/10.1037/0096-3445.123.2.178

Goldstone, R. L., Lippa, Y., & Shiffrin, R. M. (2001). Altering object representations through category learning. *Cognition, 78*(1), 27-43. https://doi.org/10.1016/S0010-0277(00)00099-8

Goldstone, R. L., & Steyvers, M. (2001). The sensitization and differentiation of dimensions during category learning. *Journal of Experimental Psychology: General, 130*(1), 116–139. https://doi.org/10.1037/0096-3445.130.1.116

Henderson, J. M., Weeks, P. A., & Hollingworth, A. (1999). The effects of semantic consistency on eye movements during complex scene viewing. *Journal of Experimental Psychology: Human Perception and Performance, 25*(1), 210–228. https://doi.org/10.1037/0096-1523.25.1.210

Hulme, C., Goetz, K., Gooch, D., Adams, J., & Snowling, M. J. (2007). Paired-associate learning, phoneme awareness, and learning to read. *Journal of Experimental Child Psychology, 96*(2), 150–166. https://doi.org/10.1016/j.jecp.2006.09.002

Jaeger, T. F. (2008). Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models*. Journal of Memory and Language, 59*(4), 434–446. https://doi.org/10.1016/j.jml.2007.11.007

Kahneman, D. (1973). *Attention and effort*. Prentice Hall.

Knowles, J. E. & Frederick, C. (2020). merTools: Tools for Analyzing Mixed Effect Regression Models. R package version 0.5.2. https://CRAN.R-project.org/package=merTools

Kurtz, K. J., Levering, K., Romero, J., Stanton, R. D., & Morris, S. N. (2013). Human learning of elemental category structures: Revising the classic result of Shepard, Hovland, and Jenkins (1961). *Journal of Experimental Psychology: Learning, Memory, and Cognition, 39*(2), 552–572. https://doi.org/10.1037/a0029178

Landau, B., Smith, L. B., & Jones, S. S. (1988). The importance of shape in early lexical learning. *Cognitive Development, 3*(3), 299–321. https://doi.org/10.1016/0885-2014(88)90014-7

Love, B. C. (2002). Comparing supervised and unsupervised category learning. *Psychonomic Bulletin & Review, 9*(4), 829–835. https://doi.org/10.3758/BF03196342

Lupyan, G. (2006). Labels facilitate learning of novel categories. In A. Cangelosi, A.D.M. Smith, & K. Smith (Eds.), *The Sixth International Conference on the Evolution of Language* (pp. 190–197). Singapore: World Scientific.

Lupyan, G. (2008a). From chair to "chair": A representational shift account of object labeling effects on memory. *Journal of Experimental Psychology: General, 137*(2), 348–369. https://doi.org/10.1037/0096-3445.137.2.348

Lupyan, G. (2008b). The conceptual grouping effect: Categories matter (and named categories matter more). *Cognition, 108*(2), 566–577. https://doi.org/10.1016/j.cognition.2008.03.009

Lupyan, G. (2009). Extracommunicative functions of language: Verbal interference causes selective categorization impairments. *Psychonomic Bulletin & Review, 16*(4), 711–718. https://doi.org/10.3758/PBR.16.4.711

Lupyan, G. (2012a). Linguistically modulated perception and cognition: The label-feedback hypothesis. *Frontiers in Cognition, 3*, 54. https://doi.org/10.3389/fpsyg.2012.00054

Lupyan, G. (2012b). What do words do? Towards a theory of language-augmented thought. In B. H. Ross (Ed.), *The psychology of learning and motivation* (Vol. 57, pp. 255–297). Waltham, MA: Academic Press. https://doi.org/10.1016/B978-0-12-394293-7.00007-8

Lupyan, G., & Casasanto, D. (2015). Meaningless words promote meaningful categorization. *Language and Cognition, 7*(2), 167-193. https://doi.org/10.1017/langcog.2014.21

Lupyan, G., Rahman, R. A., Boroditsky, L., & Clark, A. (2020). Effects of language on visual perception. *Trends in Cognitive Sciences, 24* (11)*,* 930–944. https://doi.org/10.1016/j.tics.2020.08.005

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. https://doi.org/10.1111/j.1467-9280.2007.02028.x

Lupyan, G., & Spivey, M. J. (2008). Perceptual processing is facilitated by ascribing meaning to novel stimuli*. Current Biology, 18*(10), R410–R412. https://doi.org/10.1016/j.cub.2008.02.073

Lupyan, G., & Spivey, M. J. (2010a). Making the invisible visible: Auditory cues facilitate visual object detection. *PLoS ONE, 5*(7), e11452. https://doi.org/10.1371/journal.pone.0011452.

Lupyan, G., & Spivey, M. J. (2010b). Redundant spoken labels facilitate perception of multiple items. *Attention, Perception, & Psychophysics, 72*(8), 2236–2253. https://doi.org/10.3758/APP.72.8.2236.

Lupyan, G., & Thompson-Schill, S. L. (2012). The evocative power of words: Activation of concepts by verbal and nonverbal means. *Journal of Experimental Psychology: General, 141*(1), 170–186. https://doi.org/10.1037/a0024904

Lupyan, G., Thompson-Schill, S. L., & Swingley, D. (2010). Conceptual penetration of visual processing*. Psychological Science, 21*(5), 682-691. https://doi.org/10.1177/0956797610366099

Lupyan, G., & Ward, E. J. (2013). Language can boost otherwise unseen objects into visual awareness. *Proceedings of the National Academy of Sciences, 110*(35), 14196-14201. https://doi.org/10.1073/pnas.1303312110

MacLeod, C. M., & Dunbar, K. (1988). Training and Stroop-like interference: evidence for a continuum of automaticity. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 14*(1), 126–135. https://doi.org/10.1037/0278-7393.14.1.126

Mathôt S,& Vilotijević A. (2022). Methods in Cognitive Pupillometry: Design, Preprocessing, and Statistical Analysis. bioRxiv; 2022. https://doi.org/10.1101/2022.02.23.481628.

Nosofsky, R. M. (1986). Attention, similarity, and the identification–categorization relationship. *Journal of Experimental Psychology: General, 115*(1), 39–57. https://doi.org/10.1037/0096-3445.115.1.39

Peirce, JW (2007) PsychoPy - Psychophysics software in Python. *Journal of Neuroscience Methods, 162*(1), 8–13, https://doi.org/10.1016/j.jneumeth.2006.11.017

Perry, L. K., & Lupyan, G. (2013). What the online manipulation of linguistic activity can tell us about language and thought*. Frontiers in Behavioral Neuroscience, 7*, 122. https://doi.org/10.3389/fnbeh.2013.00122

Perry, L. K., & Lupyan, G. (2014). The role of language in multi-dimensional categorization: Evidence from transcranial direct current stimulation and exposure to verbal labels. *Brain and Language, 135*, 66–72. https://doi.org/10.1016/j.bandl.2014.05.005

Perry, L. K., & Lupyan, G. (2016). Recognising a zebra from its stripes and the stripes from "zebra": the role of verbal labels in selecting category relevant information. *Language, Cognition and Neuroscience*. Advanced online publication. https://doi.org/10.1080/23273798.2016.1154974

Poldrack, R. A., Clark, J., Paré-Blagoev, E. J., Shohamy, D., Creso, J., Myers, C. E., & Gluck, M. A. (2001). Interactive memory systems in the human brain. *Nature, 414*(6863), 546–550. https://doi.org/10.1038/35107080

Pothos, E. M., Edwards, D. J., & Perlman, A. (2011). Supervised versus unsupervised categorization: Two sides of the same coin? *The Quarterly Journal of Experimental Psychology, 64*(9), 1692–1713. https://doi.org/10.1080/17470218.2011.554990

Pothos, E. M., & Reppa, I. (2014). The fickle nature of similarity change as a result of categorization. *The Quarterly Journal of Experimental Psychology, 67*(12), 2425–2438. https://doi.org/10.1080/17470218.2014.931977

Protopapas, A., Tzakosta, M., Chalamandaris, A., & Tsiakoulis, P. (2012). IPLR: An online resource for Greek word-level and sublexical information. *Language Resources and Evaluation, 46*(3), 449–459. https://doi.org/10.1007/s10579-010-9130-z

R Core Team (2021). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. Retrieved from https://www.R-project.org/.

Regier , T , Kay , P. , Gilbert , A. L. , & Ivry , R. B. (2010). Language and thought: Which side are you on, anyway? In B. Malt & P. Wolff (Eds.), *Words and the mind: How words capture human experience* (pp. 165–182) . Oxford, England : Oxford University Press

Rehder, B., & Hoffman, A. B. (2005a). Eyetracking and selective attention in category learning. *Cognitive Psychology, 51*(1), 1–41. https://doi.org/10.1016/j.cogpsych.2004.11.001

Rehder, B., & Hoffman, A. B. (2005b). Thirty-something categorization results explained: Selective attention, eyetracking, and models of category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 31*(5), 811–829. https://doi.org/10.1037/0278-7393.31.5.811

Schyns, P. G., Goldstone, R. L., & Thibaut, J.-P. (1998). The development of features in object concepts. *Behavioral and Brain Sciences, 21*(1), 1–54. https://doi.org/10.1017/S0140525X98000107

Schyns, P. G., & Rodet, L. (1997). Categorization creates functional features. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 23*(3), 681–696. https://doi.org/10.1037/0278-7393.23.3.681

Tolins, J., & Colunga, E. (2015). How words anchor categorization: conceptual flexibility with labeled and unlabeled categories‡. *Language and Cognition, 7*(2), 219–238. https://doi.org/10.1017/langcog.2014.26

Vanderplas, J. M., & Garvin, E. A. (1959). The association value of random shapes. *Journal of Experimental Psychology, 57*(3), 147–154. https://doi.org/10.1037/h0048723

Van Casteren, M., & Davis, M. H. (2006). Mix, a program for pseudorandomization. *Behavior Research Methods, 38*(4), 584–589. https://doi.org/10.3758/BF03193889

Van Gulick, A. E., & Gauthier, I. (2014). The perceptual effects of learning object categories that predict perceptual goals. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 40*(5), 1307–1320. https://doi.org/10.1037/a0036822

Van Rij, J., Wieling, M., Baayen, R., & van Rijn, H. (2020). itsadug: Interpreting Time Series and Autocorrelated Data Using GAMMs. R package version 2.4.

Võ, M. L. H., & Henderson, J. M. (2009). Does gravity matter? Effects of semantic and syntactic inconsistencies on the allocation of attention during scene perception. *Journal of Vision, 9*(3), 24–24. https://doi.org/10.1167/9.3.24.

Waxman, S. R., & Markow, D. B. (1995). Words as invitations to form categories: Evidence from 12-to 13-month-old infants. *Cognitive Psychology, 29*(3), 257–302. https://doi.org/10.1006/cogp.1995.1016

Wood, S. N. (2011). Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 73*(1), 3–36. https://doi.org/10.1111/j.1467-9868.2010.00749.x

Yoshida, H., & Smith, L.B. (2005). Linguistic cues enhance the learning of perceptual cues. *Psychological Science, 16*, 90–95. https://doi.org/10.1111/j.0956-7976.2005.00787.x