Journal of Memory and Language 119 (2021) 104242

Contents lists available at ScienceDirect

Journal of Memory and Language

journal homepage: www.elsevier.com/locate/jml

Categorical distinctiveness constrains the labeling benefit in visual working memory

Alessandra S. Souza [a,b,*], Clara Overkott [b], Marta Matyja [c]

[a] Faculty of Psychology and Education Sciences, University of Porto, Portugal
[b] Department of Psychology, University of Zurich, Switzerland
[c] Faculty of Applied Linguistics, University of Warsaw, Poland

ARTICLE INFO

Keywords:
Categorization
Color
Distinctiveness
Labeling
Shapes
Visual working memory

ABSTRACT

Describing our visual experiences improves their retention in visual working memory, yielding a labeling benefit. Labels vary, however, in categorical distinctiveness: they can be applied broadly or narrowly to categorize stimuli. Does categorical distinctiveness constrain the labeling benefit? Here, we varied the number of terms used to label continuously varying colors (Experiment 1) and shapes (Experiment 2). Participants memorized four items, and later recalled them using a continuous color or shape wheel. During study, participants articulated *"bababa"* or labeled the items with two, four, or their preferred term. Recall error decreased with increases in the number of labels. Mixture modeling showed that labeling increased the probability of recall. Memory precision, however, varied with categorical distinctiveness: broad labels reduced precision, whereas categorically distinct labels increased precision compared to no-labels. In sum, in-the-moment labeling activates categorical knowledge that facilitates the storage of visual details. Data and analysis scripts are available at: https://osf.io/mqg4k/

Introduction

"*A picture is worth a thousand words*". This English idiom illustrates the common sense that our visual experience is far richer than usually conveyed by language. Notwithstanding, it is commonplace to describe our visual experiences as they unfold in front of us or as we recollect them in our minds. Verbal descriptions (hereafter labels) can vary in many parameters, including the specificity with which they identify the visual stimulus, what we will refer to here as *categorical distinctiveness* (Murdock, 1960). For example, one could describe the color of a piece of clothing as "vibrant", or use a more specific term, e.g., "pink". The term "vibrant" can be applied broadly to categorize many colors, whereas "pink" is more narrowly applied over the same space. Accordingly, "pink" can differentiate between more colors than "vibrant". The present study assessed how the categorical distinctiveness of labels affects the quantity and quality of information stored in visual working memory.

In the following sections, we will define working memory and the role categorization plays in this memory system. Next, we will discuss the relation between categorization and verbal labeling, and the hypotheses that have been raised regarding the labeling effect on visual memory. Finally, we will delineate hypotheses regarding the role of categorical distinctiveness in visual working memory.

Visual working memory and categorization

Visual working memory keeps visual representations accessible for ongoing cognitive processing. In a prototypical visual working memory task, participants store the precise feature values of a set of stimuli (e.g., their colors), and reproduce the feature of a tested item using a continuous scale, for example a continuous color wheel (Prinzmetal et al., 1998; Wilken & Ma, 2004; Zhang & Luck, 2008). Responses in this task can be modeled to estimate how much information was accessible in working memory, and the precision with this information was stored using the so-called *mixture models* (Bays et al., 2009; Zhang & Luck, 2008). Typically, the quantity and quality of visual working memory representations decreases with increasing memory load (Luck & Vogel, 2013).

Critically, research has consistently demonstrated that categorical knowledge about visual features affects both perception (Athanasopoulos et al., 2011; Franklin et al., 2008; Hanley & Roberson, 2011; Roberson & Davidoff, 2000; Thierry et al., 2009; Winawer et al., 2007) and memory over the short-term and long-term (Bae et al., 2014; Boynton et al., 1989; Persaud & Hemmer, 2016; Uchikawa & Shinoda,

* Corresponding author at: Faculty of Psychology and Educational Sciences, University of Porto, Rua Alfredo Allen s/n, 4200-315 Porto, Portugal.
E-mail addresses: alessandra@fpce.up.pt (A.S. Souza), c.overkott@psychologie.uzh.ch (C. Overkott), martamatyja1994@gmail.com (M. Matyja).

https://doi.org/10.1016/j.jml.2021.104242
Received 20 April 2020; Received in revised form 9 March 2021; Accepted 13 March 2021
Available online 5 April 2021

1996). Accordingly, more recently researchers have extended the mixture models used to assess performance in visual working memory tasks to also account for the influence of categorical knowledge. In essence, these models allow some responses to be based on categorical representations (e.g., red vs. green), whereas other responses are assumed to be based on continuous (fine-grained) representations of the exact feature value experienced (Bae et al., 2015; Cibelli et al., 2016; Hardman et al., 2017; Pratte et al., 2017). These studies have shown that a substantial amount of responses in visual working memory tasks are influenced by the prior knowledge of the participants, as reflected in well-learned feature categories.

Categorization and verbal labeling

Implicitly, the use of categorical knowledge in visual working memory tasks has been assumed to be a by-product of verbal labeling (Cibelli et al., 2016; Hardman et al., 2017) but the reliance on labeling has been hardly directly manipulated. For example, in the study of Donkin, Nosofsky, Gold, and Shiffrin (2015), participants were asked to store the precise color of a single dot presented for 0.1, 0.5, or 2 s. In most trials, offset of the stimulus was followed by a delay of variable duration, after which visual working memory was tested with a color wheel. In other trials, participants labeled the color and color reproduction based on the label was tested three trials later. Longer study durations yielded more precise visual memory of the stimulus, as well as more precise responding in labeling trials. They modeled the visual working memory data incorporating several sources of information: the perceptual information from the stimulus, the precision of verbal labels, and random guessing. All of these factors proved necessary to account for their data. Although labeling precision was directly assessed in this study, they did not manipulate reliance on the labels and hence the study could not inform whether labeling changed how information was stored in visual working memory.

There are multiple ways in which labeling could alter the storage of visual information in memory. Several hypotheses of the labeling effect have been raised with regards to investigations of episodic long-term memory. We will briefly review these hypotheses below.

Verbal recoding

The *verbal recoding* (Souza & Skóra, 2017) or *verbal overshadowing* (Alogna et al., 2014; Schooler & Engstler-Schooler, 1990) hypothesis states that labeling promotes the storage of a verbal representation at the expense of the visual one. Storage of the label *"green"* at the expense of the particular greenish hue presented for study should lead to a large loss of precision in recalling this feature from memory. For example, in the classical study by Schooler and Engstler-Schooler (1990), worse memory for a face was observed when participants were requested to describe the stimulus during the retention interval. Several studies have observed costs of labeling, although facilitation has been observed in some conditions (for a *meta*-analysis and overview of the verbal overshadowing effect see Meissner et al., 2008).

Memory distortion

Lupyan (2008) proposed the shift-to-prototype hypothesis which states that labeling causes visual representations to more strongly drift towards the category prototype activated by the label. This hypothesis was based on the observation that asking people to classify objects as belonging to one out of two categories (e.g., lamps vs. chairs) impaired episodic visual long-term memory for the studied exemplars compared to memory of the same objects studied under a preference rating instruction. The idea here is that top-down categorical information can distort the storage of perceptual information provided by the perceptual input.

Categorical information is known to guide memory responses irrespective of labeling (Crawford et al., 2000; Huttenlocher et al., 1991) and the ability to form conceptual classes is also displayed by non-human animals which do not have language to assist in this process (Zentall et al., 2008). However, labels have been found to activate categorical information more strongly than would have occurred only by the presentation of an exemplar of the category (Boutonnet & Lupyan, 2015; Edmiston & Lupyan, 2015; Forder & Lupyan, 2019; Lupyan & Thompson-Schill, 2012). In this way, online labeling of a stimulus during study would accentuate the memory distortion produced by prior experience with categories. This is in agreement with the results reported by Carmichael et al. (1932) in which ambiguous line-drawings were presented to participants and these were paired with one of two labels. Free drawing of the images from memory was biased by the labels. For example, when a drawing was paired with the term "moon", it was sketched later more similarly to a moon than when the same drawing was paired with the label "C".

In sum, this hypothesis predicts an increase in the ability to discriminate stimuli from different categories, but reduced ability to make within-category discriminations. Contradicting this hypothesis, Forder and Lupyan (2019) showed that in-the-moment labeling increased between-category as well as within-category discrimination of colors.

Dual-Trace (Visual + Verbal)

Instead of replacing the visual trace, labeling could add a verbal representation to the visual one, thereby creating two sources of information: a continuous visual representation and a verbal one containing categorical information. Both sources of information may guide recall or one of them may dominate depending on the situation, in line with the dual-coding theory (Pavio, 1991). This hypothesis guided the modeling implemented by Donkin et al. (2015): verbal labeling was considered as an additional source of information. Other studies have shown that labeling may impair performance if the verbal label dominates. Labeling can be inconsequential if the context for the visual item is reinstated or if it occurs during the study phase instead of the retention interval (Brandimonte et al., 1992, 1997; Brown et al., 2014). These results suggest that people had two independent sources of information in mind: the visual trace and the verbal trace.

Distinctiveness

The label could serve as an additional retrieval cue to the visual trace. In this scenario, labeling helps to the extent that it distinguishes between items in memory. This account explains the labeling costs observed by Lupyan (2008): the lack of distinctiveness of the labels used in this study (two labels for several exemplars of the same category) would reduce their distinctiveness. In contrast, Richler et al. (2013) varied whether studied exemplars were from two categories or unique categories. Labeling only yielded worse long-term memory than preference rating in the two-category but not in the unique-category condition. This was not simply due to an increase in categorical memory because labeling improved rejection of within-category lures, suggesting that label distinctiveness impacts the storage of visual details (see also Bartlett, Till, & Fields, 1980).

Categorical visual long-term memory

Yet another possibility is that labels activate a visual trace of the category in long-term memory. Participants would then have two visual traces to rely on, with the visual trace of the category serving to enhance the perception of the specific value of the presented item or to protect it from interference. The *label-feedback* hypothesis (Lupyan, 2012a) states that saying (or hearing) a label such as *"green"* would transiently activate visual features related to green that set it apart from other categories. This could then serve to sharpen the perception of a greenish hue presented for study, arguably protecting it from forgetting, increasing its fidelity, or facilitating its short-term consolidation (Ricker, 2015). As we reviewed above, verbal labeling has been found to be a more effective means to activate categorical information than properties of category exemplars (Boutonnet & Lupyan, 2015; Edmiston & Lupyan, 2015;

Forder & Lupyan, 2019; Lupyan & Thompson-Schill, 2012; Lupyan & Ward, 2013). For example, the sound of a barking dog is less efficient to cue the category of dog than the word "dog" (Edmiston & Lupyan, 2015). In line with this possibility, Bower et al. (1975) observed an effect of presenting abstract drawings (droodles) with and without an associated verbal interpretation that gave it meaning. When a meaningful interpretation was provided, memory was augmented and participants could recall more drawings compared to a baseline condition with no interpretations.

Labeling in visual working memory

The hypotheses listed above were raised and tested mainly with regards to long-term retention of information. The impact of labeling in visual working memory has been much less investigated. This is mostly because visual working memory studies often use a set-up that discourages labeling from occurring. First, all of the visual stimuli are presented in a one-shot display; second, the presentation duration is usually only a few hundred milliseconds; and third, memory is tested shortly after (typically 1 s). This strongly reduces the opportunities for labeling to occur. Accordingly, performance in change detection tasks (that only require recognition of a change in the display) does not show effects of blocking verbal labeling, as tested by imposing verbal memory loads (Vogel et al., 2001) or the use of an articulatory suppression procedure in which irrelevant syllabi are articulated continuously throughout study and memory retention (Morey & Cowan, 2004, 2005; Sense et al., 2016).

To assess the impact of providing opportunities to label the visual memoranda, Souza and Skóra (2017) used a continuous color reproduction task in which the memoranda were presented sequentially with sufficient inter-item interval for the generation of a label. They compared an articulatory suppression condition that prevented in-the-moment labeling (i.e., saying "bababa" aloud) with a condition in which colors were overtly labeled by the participants. The labels were recorded and later coded by the experimenter, revealing that a set of seven color terms – that is, red, orange, yellow, green, blue, purple, and pink – were used in most trials. As expected, color labeling increased the quantity of categorical representations in visual working memory. Unexpectedly, however, labeling also increased the quality (and sometimes the quantity) of the continuous representations stored in working memory, indicating that representations of the labeled information were stored with higher fidelity. They also observed that other types of labeling that lacked categorical information, for example, labeling the order in which the items were presented (e.g., "first", "second", etc.) was not beneficial; instead it produced similar performance as in the suppression condition. Souza and Skóra (2017) reasoned that this effect was due to the color labels providing categorical information about the relevant memory feature thereby helping to protect the continuous representation of this feature from interference from the other memory items or from the test situation. In contrast, labeling the order of presentation of the memoranda was not beneficial although it provided a distinct label to each item in the memory array. They argued this was the case because these labels lacked categorical information. This pattern challenges the distinctiveness hypotheses discussed above, but it is the pattern predicted by the categorical visual long-term memory hypothesis.

Forsberg et al. (2020) replicated the design of Souza and Skóra (2017) in a sample of younger and older adults. Labeling also improved performance compared to a suppression condition, but the source of the benefit was different between younger and older adults. In young adults, color labeling afforded the storage of more visual details as well as more categorical information, replicating Souza and Skóra (2017). Conversely, older adults only showed a benefit in categorical memory, but not in continuous memory with labeling. These results show that the benefits of labeling may be subject to age-related cognitive decline. Unfortunately, in this study the labeling behavior of the participants was not recorded and hence there was no information regarding the number and variety of labels used by the younger and older participants. It remains open the question of whether the quality of the labels used by the participants could explain the differential impact of labeling with aging. As we will argue below, how people categorize and label the visual stimuli can have a profound impact on the retention of this information in visual working memory, and whether the labels increase memory precision.

Categorical distinctiveness of the Labels

Souza and Skóra (2017) showed that simply using distinct labels for each item in the memory array (i.e., position labels to remember colors) did not yield a labeling benefit, but using labels that carry categorical information does (i.e., color labels to remember colors). This is inconsistent with a simple distinctiveness hypothesis of the labeling benefit. However, it does not rule out a role of distinctiveness in terms of the categories activated by the labels in long-term memory. As previously pointed out, most of the generated labels in the color labeling condition of Souza and Skóra (2017) comprised one of seven basic color terms. Although research has shown that some color categories are already present in infancy (Skelton et al., 2017), color categories receive much influence from cultural practices that shape color distinctions (Regier et al., 2007). Hence it is reasonable to assume that the choice of labels is related to their categorical distinctiveness in a culturally defined space, i.e., how much they differentiate between items in the feature space. This differentiation is however not fixed, and this could also pose a constrain in the benefits one can draw from labeling. If fewer terms are used to categorize the same set of colors, then the labels would be less distinct, and the beneficial effect of the labels could be reduced. Yet, to the best of our knowledge, no study manipulated the categorical distinctiveness of labels to trace its impact on visual working memory storage.

Some finding from the episodic visual long-term memory literature suggest that the number of categories along which items have to be categorized/labeled may define whether labels produce costs or are inconsequential. For example, in the study by Lupyan (2008) worse long-term memory was observed when exemplars were labeled as belonging to two categories (chairs or lamps) compared to a preference rating condition. Richler et al. (2013) varied whether studied exemplars were from two categories or unique categories. Labeling only yielded worse long-term memory than preference rating in the two-category but not in the unique-category condition (see also Bartlett et al., 1980).

In the studies listed above the number of categories varied with the type of memoranda studied, namely when fewer terms were used to label the stimuli, these stimuli were also sampled from a small set of categories. This limits the conclusions one can draw regarding the online impact of categorization on memory. Here we assume that labeling involves the activation of learned categories in long-term memory in line with the label-feedback hypothesis (for reviews see Lupyan, 2012a, 2012b). Critically, the same set of stimuli can be categorized or labeled differently depending on the task goals, thereby producing different levels of abstraction and distinctiveness (Pansky & Koriat, 2004), and this can provide a flexible and task-dependent modulation of memory. Take, for example, the visual domain of color. Recent research has shown that hearing color labels can warp perception of colors increasing their discriminability even when these labels present redundant information (Forder & Lupyan, 2019). This effect was observed both for between-category comparisons, as well as within-categories comparisons (i.e., discriminating between more prototypical and less prototypical colors in the same category). Little is known, however, regarding what happens when people are faced with labels that are less categorically distinct (e.g., warm vs. cold colors). If labels have a flexible and online (i.e., in the moment) influence on perception, attention, and memory as predicted by the label feedback hypothesis, categorical effects should change depending on which category the label activates.

The present study

Here, we provide the first direct assessment of the impact of the categorical distinctiveness of labels on the quantity and quality of visual working memory. To manipulate categorical distinctiveness, we experimentally varied the number of labels (2 or 4) used to describe continuously-varying colors (Experiment 1) and continuously-varying shapes (Experiment 2) memorized for a visual working memory task. We also implemented two other conditions, one in which labeling was hindered with the imposition of articulatory suppression (repeatedly say "ba ba ba" aloud, hereafter the AS condition), and one in which participants were free to label the items as they wished (Free Labels condition). These conditions replicate the ones used to assess the labeling effect by Souza and Skóra (2017). The AS condition serves as a baseline in which the usage of labeling is hindered. This permits the assessment of categorical effects that occur irrespective of labeling. The Free Labels condition allows one to assess the maximum benefit of using as many distinct labels as participants can think of based on their prior learning experience. The 2-Labels and 4-Labels conditions, conversely, allow us to assess whether labeling can have a flexible and task dependent effect on memory depending on how the visual objects are categorized in the moment.

With these manipulations, our study aims to provide a clearer picture of the effects of label categorical distinctiveness than achieved in prior studies. For example, in Richler et al. (2013) label distinctiveness was confounded with the type of memoranda: the low-distinctiveness label condition involved studying multiple exemplars from the same category, whereas the high-distinctiveness label condition involved studying unique exemplars from different categories. Accordingly, these conditions could not be directly compared, and their impact was assessed in relation to other forms of encoding instructions (e.g., preference rating). One problem with comparing labeling to other forms of encoding instruction is that there is no means to know whether labeling was not beneficial or whether these other conditions were just more beneficial for memory than labeling. For example, Blanco and Gurenckis (2013) showed that preference rating yields good memory because of the larger distinctiveness and deep processing it affords. When they compared labeling to another encoding instruction (orientation of the object), labeling yielded comparable levels of performance. The present study overcomes this difficulty by comparing conditions that promote and hinder verbal labeling directly (by requiring different types of overt labeling vs. articulatory suppression), and we compared the impact of these manipulations upon memory of the same set of visual stimuli, removing confounds of differences in memorability between stimulus sets. Lastly, our task also allows one to estimate the quantity and quality of memory representations using modeling, and hence to address the hypothesis that labeling impacts the storage of visual details.

Our predictions were as follows. First, we aimed to replicate the labeling benefit observed by Souza and Skóra (2017) when contrasting the AS vs. the Free Labels conditions. Overtly labeling colors or shapes with their preferred term in the Free Labels condition should allow participants to store more precise representations of the memory items. This has been taken as evidence that labeling has an online effect on memory. According to the label-feedback hypothesis, the act of labeling highlights typical or diagnostic properties of the category while irrelevant properties can be abstracted (Lupyan, 2012b). If this is indeed the case, varying the categorical distinctiveness of the labels between the 2- and 4-Labels conditions should change the categories activated in visual long-term memory, and hence alter the prior these categories offer to ground the noisy information that will be stored in memory. When the labels are less distinct (such as in the 2-Labels condition), they will highlight a broad range of features, providing an imprecise prior for encoding the information and leading to lower memory fidelity. When the labels are more distinct (e.g., 4-Labels condition), the categories are narrower and the more precise prior would lead to more precise representations. This would show that the tradeoff in memory precision induced by labeling depends on the categorical distinctiveness of the labels.

Experiment 1

In Experiment 1, we used the color reproduction task employed by Souza and Skóra (2017): Participants remembered four colored dots presented sequentially and, at test, they reproduced each color using a continuous color wheel. Across four different within-subject conditions, we varied the reliance on labels to categorize the colors during study. In the AS condition, labeling was prevented with an articulatory suppression procedure. In the Free Labels condition, participants could label the colors with as many terms as they wished. In the 2-Labels and 4-Labels conditions, we sectioned the color wheel into two and four parts, and we let participants select a label to refer to these sections. We then trained them in consistently applying these terms to categorize colors in each of the sections. Afterwards, participants had to use these labels during the following color memory task.

If labeling has an online effect in memory as predicted by the visual long-term memory hypothesis of Souza and Skóra (2017) which was derived from the label-feedback hypothesis of Lupyan (2012a, 2012b), then we should observe that the benefits of the labels will vary with their categorical distinctiveness.

Methods

Participants

Forty-eight students (39 women; $M = 22.75$ years old) from the University of Zurich completed two 1-hour sessions in exchange of course credit or 30 Swiss francs (ca. 30 US dollars). The experiment consisted of a within-subjects design in which all participants were exposed to all of our four experimental conditions. We based our sample-size decision on the number needed to counterbalanced the order of our experimental conditions across participants (16 possible orders were created – see constrains below), and we replicated this counterbalancing three times. Given that we used Bayesian inference to assess evidence for changes in recall performance, we collected a sufficiently large sample to yield strong evidence (Bayes Factor, $BF > 10$) for changes in recall error across conditions. In Bayesian statistics it is not a problem to increase sample-size after seeing the data to obtain stronger evidence because this does not unduly inflate false positives. This is because evidence is computed both in favor and against the presence of an effect (Rouder, 2014; Schönbrodt et al., 2017).

Four participants were excluded due to non-compliance with the verbalization instructions, yielding a final $N = 44$. All remaining participants complied with the labeling instructions and showed proper labeling behavior (described below).

For all experiments described here, participants were German-speaking individuals from the Zurich area of Switzerland that reported normal or correct-to-normal vision, and no color blindness. Participants signed an informed consent form prior the study, and were debriefed at the end. The study followed the ethical guidelines of the institutional ethics review board, and it did not require special ethical approval.

Procedure

The main experiment consisted of a continuous color reproduction task. Additionally, participants completed two pre-training tasks to establish accurate labeling behavior in the 2-Labels and 4-Labels conditions. The experimental task was programmed in Matlab using the Psychophysics toolbox (Brainard, 1997; Pelli, 1997). All instructions and the labeling behavior of the participants occurred in German.

Continuous color reproduction task. In the beginning of every trial (see Fig. 1A), a grey background (RGB 128 128 128) with four dark grey discs (RGB 112 112 112; 35 pixels radius) evenly spaced on an imaginary

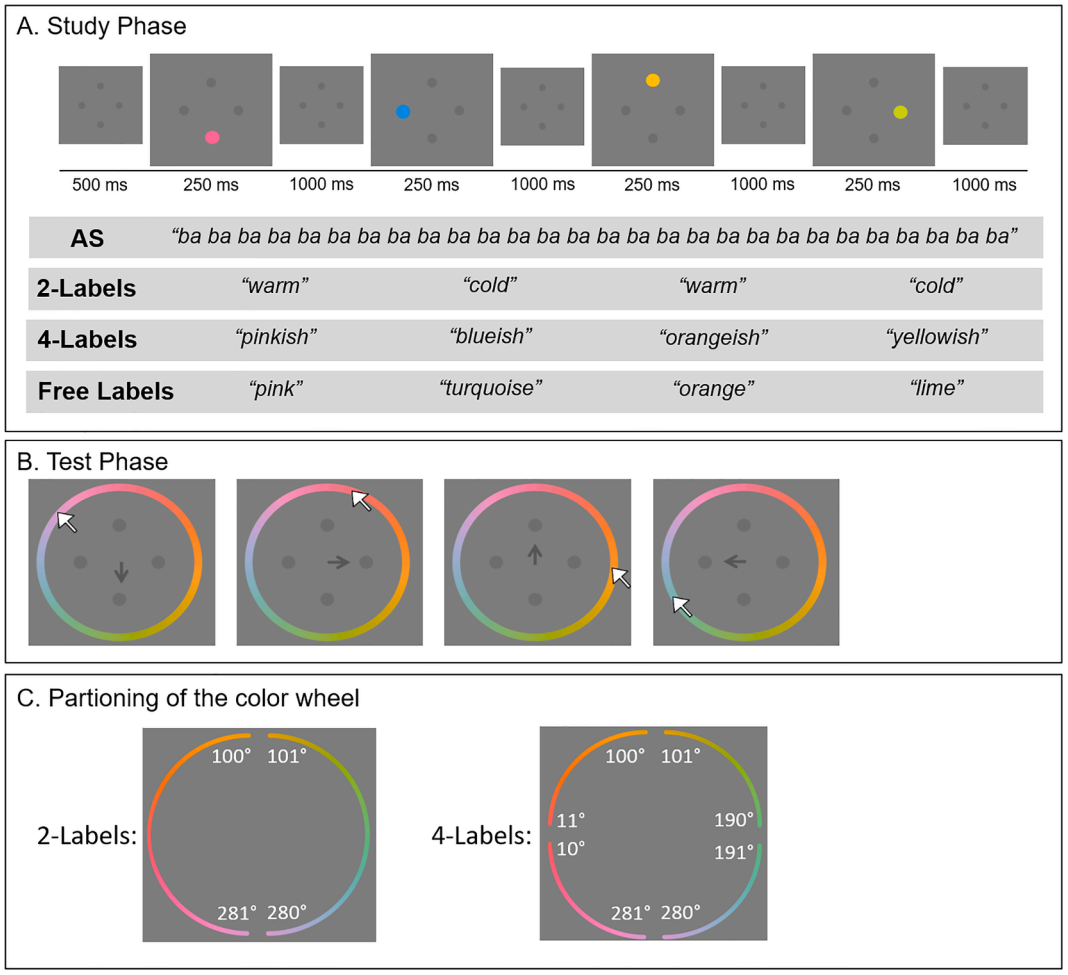

Fig. 1. Experimental Design of Experiment 1. Note. Panel A illustrates the flow of events in the study phase and the verbalization requirements in each experimental condition in Experiment 1. Note that labels are illustrative and were individually determined. Panel B illustrates the flow of events in the test phase. The dark grey arrow indicated the randomly selected item to be reproduced. Panel C shows the division of the color wheel presented to participants. Participants were asked to type a label to refer to each of the sections. Note that the angles in the wheel printed in this illustration were not visible to participants, and are presented here for informative purposes only. Angles refer to the un-rotated color wheel.

circle (200 pixels radius) centered in the middle of the screen was presented for 500 ms. Next, a color was presented at one disc at a time for 250 ms, followed by an inter-stimulus interval of 1000 ms. Colors were randomly sampled from 360 values evenly distributed along a circle in the CIELAB color space with L = 70, a = 20, b = 38, and radius = 60 (Zhang & Luck, 2008). The first stimulus position was randomly selected from the four locations, and the subsequent ones followed in clockwise fashion.

Next, memory for all four colors was tested in random order (see Fig. 1B). A color wheel (randomly rotated from trial-to-trial) was shown around the four grey dots and a central dark grey arrow indicated the stimulus to be reproduced on it. The arrow pointed to one location at a time, in random order, until all stimuli were tested. Participants moved the mouse around the wheel to adjust the color of the tested item, and they clicked with the mouse to confirm their response.

Participants completed four blocks of 50 trials, each block consisting of one labeling condition (see Fig. 1A). In the AS condition, participants repeated *"bababa"* aloud throughout the study phase, thereby preventing labeling. In the Free-Label condition, participants freely labeled the colors with any term they wanted. These conditions replicate the ones used by Souza and Skóra (2017) in demonstrating a labeling benefit in visual working memory. Our experimental manipulation of labeling distinctiveness was implemented in the 2-Labels and 4-Labels conditions in which participants labeled the colors using only two and four terms (individually defined, see pre-training below). Trial-by-trial verbalizations were recorded, and coded offline by the experimenter.

Each session consisted of two blocks. The 2-Labels and 4-Labels blocks did not occur in the same session due to time constrains related to the pre-training phase. Block order within the session and session order were counterbalanced across participants.

Label Pre-Training. The 2-Labels and 4-Labels blocks were preceded by a pre-training phase. First, participants were shown the color wheel split in two 180° sections for the 2-Labels or four 90° sections for the 4-Labels block, respectively (see Fig. 1C). Participants typed a label to each section. Next, participants trained applying these labels to the colors on the wheel. A color appeared for 250 ms, followed by a 1000 ms labeling window. Afterwards, the pre-specified label was printed onscreen, and participants pressed the right or left arrow-key to indicate whether they labeled the color correctly or incorrectly, respectively (self-scoring). Training comprised, at least, 100 trials. It continued until a minimum self-reported accuracy of 80% over the last 50 trials, or until 360 trials were completed. If the criterion was not met, the experiment stopped (and the participant was excluded). All participants learned the labels.

Data analysis

Verbal output

For every memory trial, we recorded the verbal response of the participants during the study phase to check with instruction compliance. Four participants were excluded because they did not label the stimuli according to the instructions in the memory phase. Furthermore, verbal responses to each stimulus in the memory array were coded to assess which labels were applied to which colors. This allowed us, for example, to calculate the proportion of times labels like "orange", "green", "yellow", and so forth were applied to each of the colors in the wheel. For each label, we then draw a line joining together the proportions of times this label was used for each color. This also allowed us to assess the consistency in applying the trained labels during the memory trials in the 2-Labels and 4-Labels conditions. Given that the labels in these conditions were individually defined, we will generally term them Label 1, Label 2, and so forth. Finally, this data also served to estimate the color categories (defined as the mean of the distribution of verbal responses over the color space) for modeling of the data (see further details below).

Memory performance

The main dependent variable in our study was the absolute distance between the reported color and the true color of the tested stimulus (hereafter recall error). This provides a model-free estimate of memory accuracy. We assessed changes in recall error with Bayesian inferential statistics using the Bayes Factor package (Morey & Rouder, 2015) implemented in R (R core team, 2017) using the anovaBF function. In this model, individual mean recall error in each condition served as the predicted variable, the labeling condition served as the predictor, and participant was treated as a random effect.

Mixture modeling

A mixture model was applied to the data to estimate the probabilities that responses were sampled from several distributions reflecting memory and guessing states. We used the hierarchical Bayesian categorical-continuous mixture model of Hardman et al. (2017), which permits the estimation of the proportion of items remembered categorically *versus* continuously. The model structure is illustrated in Fig. 2. In essence, this model assumes that representations in memory are either categorical (some canonical values) or continuous (fine-grained detail about the studied feature). Responses informed by categorical information cluster around some canonical values (see Fig. 2a). Responses informed by continuous information vary linearly with the studied feature (see Fig. 2b). The fidelity of continuous information can vary, which is captured by the model parameter called continuous imprecision[1]. This is reflected by the width of the diagonal line in Fig. 2b. Responses not informed by memory are assumed to be guessing. Guessing can be categorical (random selection of one of the color categories irrespective of the studied color value; see Fig. 2c) or uniformly distributed (continuous guessing; Fig. 2d). Response distributions in the task are assumed to reflect a mixture of these four states. Accordingly, the mixture model aims at estimating the contribution of each of these distributions to generating the data modeled.

Note that this model does not include a parameter to deal with the possibility of confusion between memory items. Given that memory items were randomly selected, any misreporting of a non-tested item will be randomly distributed in relation to the value of the currently tested item and will be accounted for by the model as guessing. This does not impact estimation of correct recalls in the model, which is the main focus of this study.

[1] This is the sigma parameter of the von Mises distribution, which has the same meaning as the imprecision parameter in the traditional mixture model proposed by Zhang & Luck (2008).

In sum, this model has three main parameters: (a) the probability of storage in memory (P^M); (b) the probability that the representation in memory was continuous (P^O) as opposed to categorical (1 - P^O); and (c) the imprecision of the continuous representation in memory (σ^O). The model was built to allow these three parameters to vary between simultaneously-modelled experimental conditions. In other words, if an experiment has two conditions A and B, and the data of both conditions are modeled simultaneously, the model will allow each condition to have their own estimated values for these three parameters. The model also includes other parameters that are freely estimated for the data, but that are not allowed to vary between simultaneously-modeled conditions: the number of categories, their center and width (i.e., standard deviation), the probability of categorical guessing (P^{AG}), how colors are assigned to categories (category selectivity, σ^S) which accounts for the possibility of classification errors, and the imprecision on the selection of the category (σ^A), which captures the fact that categorical responses can deviate slightly from the category center (see width of the categorical bands in Fig. 2a). Concretely, this means that our hypothetical conditions A and B would vary in P^M, P^O and σ^O, but the model would estimate a single value for both of them reflecting the number of categories, their center and width, categorical guessing, and so forth. In other words, any differences in performance between these conditions would have to be captured in the three main parameter with condition effects and nowhere else.

Given that the model set-up does not allow simultaneously-modelled conditions to vary in parameters that reflect categorical bias (e.g., number of categories, their center and imprecision), and this was the main factor along which our conditions varied, we modeled the data of each condition separately. This way, the model estimated the contributions of categorical biases independently for each condition, thereby allowing all model parameters to capture the variation in categorical effects induced by our labeling manipulations. We fitted two types of models to the data of each condition, one in which we let the model freely estimate the number of categories and their centers, and one in which we fixed the number of color categories and their center based on the labeling behavior recorded during the memory trials.

For the Fixed Categories modeling, we took two assumptions. First, we assumed that the categories used in the Free-Labels condition reflect general prior experience with categorizing colors, and that these categories would guide memory in the presence and absence of labeling. Hence the categorical effects would be similar in the AS condition and the Free Labels condition, and we used the values estimated from the labeling behavior in the Free Labels condition to fix the categories in the AS condition. This reflects the assumption that labeling would activate categorical information to a larger degree, but without labeling categorical information would also be activated by the stimulus. This is in line with the modeling implemented by Souza and Skóra (2017), and does match the overall categorical responding observed in the raw data across these conditions (see Fig. 4). Second, to make a strong test of the online effect produced by the manipulation of labeling implemented here, we fixed the maximum number of categories to two and four in the 2-Labels and 4-Labels conditions, respectively. We also defined the center of each wheel section as the respective center of their labeling categories.

To evaluate changes in model parameters as a function of experimental condition, we considered the overlap on the posterior distribution of the parameters. The credible values in the posterior are those within the 95% highest density interval (HDI). When the HDI of two posteriors do not overlap, or the difference between the posteriors do not include zero, their difference is credible (Kruschke, 2013).

There are two model variants in the CatContModel. The between-item model assumes that each response is based on either a categorical or a continuous representation. The within-item model assumes that responses are based on a weighted combination of a continuous and a categorical representation of each item. Both Hardman et al. (2017) and Souza and Skóra (2017) found that the between-item model had a better

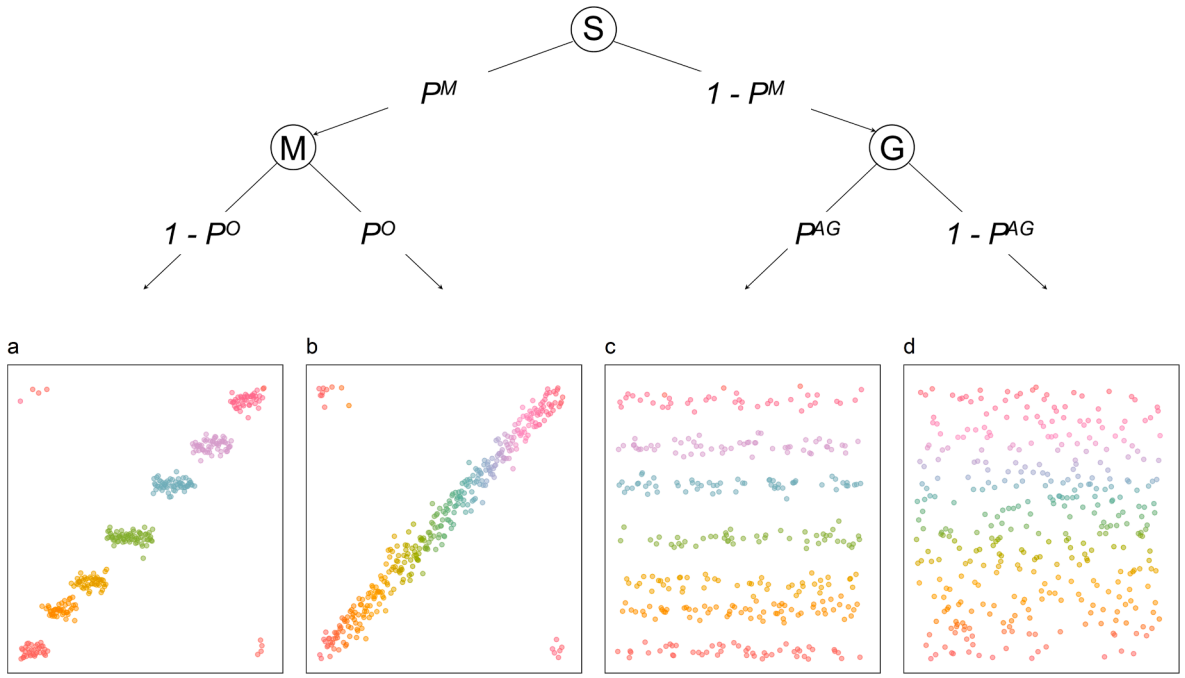

Fig. 2. Multinomial Process Tree Illustrating the Categorical-Continuous Model of Hardman et al. (2017). Note. For all scatterplots, the x-axis represents the studied color-hue and the y-axis the response hue. Each panel shows the predicted data pattern that the model parameter accounts for. Panel a. Categorical memory: for a range of studied hues, the same categorical response is provided ("red"). The width of the categorical bands reflects categorical imprecision in reporting the category. Panel b. Continuous memory: responses vary linearly with the studied hue thereby generating a continuous diagonal line. The width of the diagonal line indicates the continuous memory imprecision: a thin diagonal represents more precise co-variation of studied and response values. Panel c. Categorical guessing: guessing is distributed over constant category bands. Panel d. Random guessing. Reprinted from Souza and Skóra (2017). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

fit to the data of delayed estimation tasks. Here we fitted the between-item model version to the data of Experiment 1 (20′000 iterations; 500 burn-in) using the CatContModel package (Hardman, 2016) implemented in R.

Materials, data, and analysis scripts for all experiments reported here can be found at the Open Science Framework: https://osf.io/mqg4k/

Results

Verbal labeling

Analysis of the trial-by-trial labeling behavior of the participants during the working memory trials is presented in Fig. 3. Fig. 3A-C present the proportion of times the labels were applied to each color on the wheel in the 2-Labels, 4-Labels, and Free Label conditions in Experiment 1. Each line tracks the frequency of usage of one label over the color space. When the line is at 1.0, it indicates that this label was used by all participants to describe that color when it was presented for study in the memory trial.

As shown in Fig. 3A and B, labeling was highly accurate in the 2-Labels and 4-Labels conditions indicating that participants complied with the instruction to label the colors during the memory study phase with the trained labels. In the Free Labels condition, seven basic color terms were used in most trials (ca. 86%), reflecting high consistency in the selection of color terms by participants. Note that in the Free Labels condition, participants could use any term they wanted, but the seven color terms depicted in Fig. 3C were the most common. Overall, there was very high agreement between participants on how to label each of the colors they were trying to memorize.

Fig. 3D presents the overall usage proportion of three classes of labels during the working memory trials. The first class is what we termed basic labels (i.e., the trained labels in the 2- and 4-Labels conditions and the seven basic color terms used in the Free Labels). The second class consists of rarer labels (e.g., olive, kiwi, gold, light blue) that we labeled as "other". The last class is NA which constitutes unclassified responses, i.e., instance in which the participant remained silent or produced an unintelligible response. As expected, overall usage of the trained labels was very high, this indicated that our procedure was successful in generating different labeling behaviors in each memory condition.

Model free indices of performance

Raw responses. Fig. 4 shows a scatterplot relating studied color hue and response hue in each experimental condition (Hardman et al., 2017). In these scatterplots, it is possible to visualize the contributions of the different sources of information the mixture model aims to account for. For example, the scattered dots within each panel in Fig. 4 are consistent with guessing, the dots that align in the diagonal line with continuous memory for the studied colors, and the vertical bands along the diagonal with categorical responding. Overall, one can infer from these plots that there was more guessing under the AS condition, and less guessing under the Free Labels condition. The pattern of categorical effects is similar between the AS and Free conditions. In the 2- and 4-Labels conditions, conversely, the categorical bands mimic the ones specified by the labels at the time of study, i.e., two and four bands, respectively.

Mean Recall Error. When we take the absolute difference between the studied and reported color value, we obtain a measure of recall error. Recall error in Experiment 1 is depicted in Fig. 5A. Recall error was higher when labeling was hindered with suppression (AS condition), and lowest in the Free Labels condition. Performance in the 2- and 4-Labels conditions remained in between these values. Across all conditions, we observe a monotonic decrease in error with increases in the number of labels used. Accordingly, a one-way repeated measures BANOVA having condition (AS, 2, 4, and Free) as predictor showed overwhelming support for an effect of condition in recall error, $BF_{10} = 1.75 \times 10^{28}$. To follow up on this effect, we contrasted adjacent levels of the condition

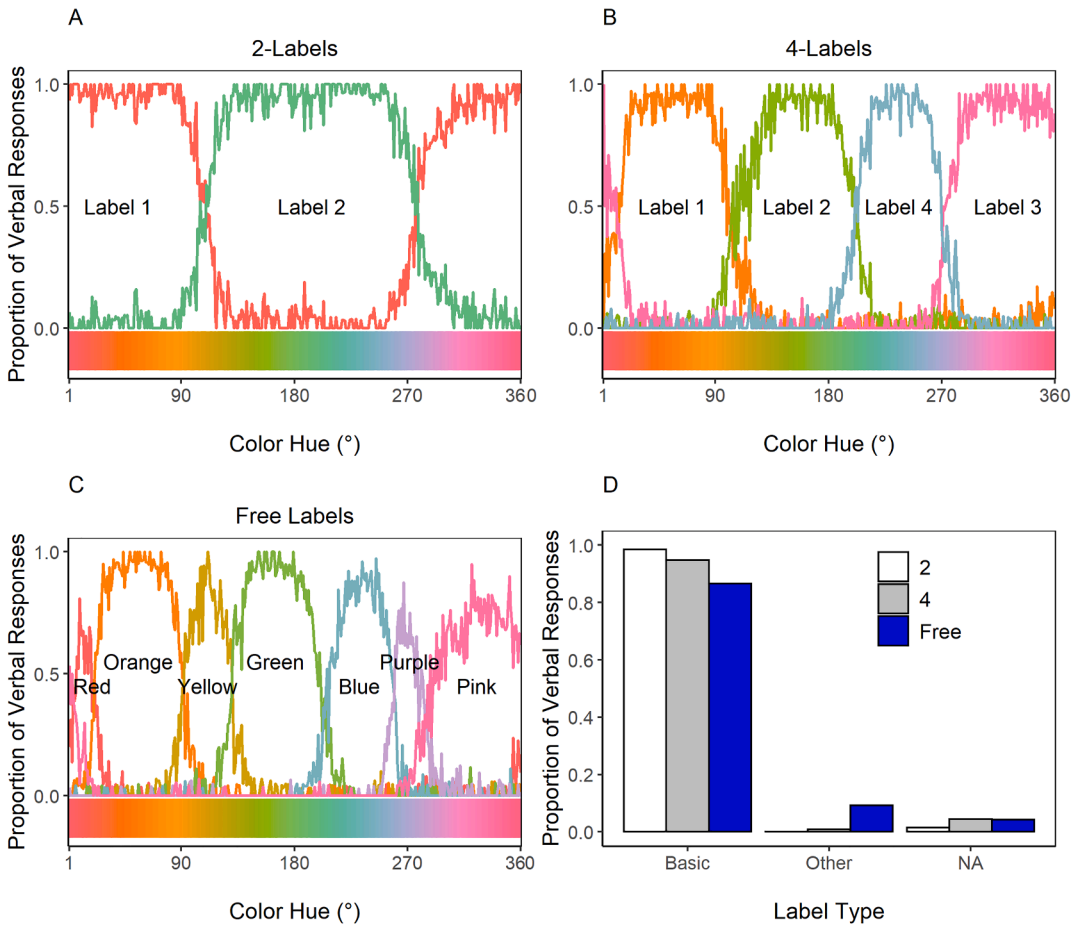

Fig. 3. Verbal Labeling Recorded During the Working Memory Trials in Experiment 1. Note. Panels A-C: Each line represents the proportion of times each label was applied to each color when it was presented for study on the working memory trials. Panel D: Overall usage proportion of the basic color terms depicted in panels A-C, other terms, or not classified responses (NA = silence or unintelligible).

variable using Bayesian *t*-tests. There was substantial evidence for a reduction in recall error between the AS and 2-Labels conditions ($BF_{10} = 32$), between the 2- and 4-Labels conditions ($BF_{10} = 4.39 \times 10^5$), and between the 4- and Free Labels conditions ($BF_{10} = 2705$).

Mixture modeling

We modeled the data either by fixing the number and center of the categories (Fixed Categories) or by allowing the model to freely estimate the number and center of the categories (Free Categories). For both types of models, we fitted them using 20,000 iterations, and discarded the first 500 iterations as burn-in. Both types of models yielded similar results, the only difference being that when the model freely estimates the categories, it is more liberal in assigning them and estimates more categorical responding than continuous responses. This, however, did not change the relative positioning of the conditions in relation to each other in either type of model. The Supplementary Materials file shows a posterior predictive check of both models and indicate that both accounted for the data well.

Categorical Information. The free-categories model provides an estimation of the number of categories that are needed to capture the most variance in performance in each condition. The estimated mean number of categories and their credible intervals were the following: AS = 7.79 [7.38, 8.20], 2-Labels = 5.53 [5.02, 6.09], 4-Labels = 6.44 [6.13, 6.77], and Free Labels = 7.97 [7.61, 8.34]. Note that the estimated number of categories was relatively high in all conditions. In our experimental conditions (i.e., 2- and 4-Labels conditions), the number of estimated categories was much larger than the number of categories we prompted participants to use. This is probably due to two factors. On the one hand, there is the issue of model flexibility. The model can potentially account for more variation by assigning large number of categories. On the other hand, the requirement to label the memoranda with a small set of terms does not wash out people's history of categorization, which may still contribute to performance. Hence the estimated values might reflect a combination of both of these factors, and should be interpreted with caution. Critical for our research question, however, is the difference in estimated categorical behavior between the conditions. The estimated number of categories did not credibly differ between the AS and Free Labels conditions (i.e., their credible intervals overlap), but values in these conditions were credibly higher than in the 2-Labels and 4-Labels conditions. This indicates that: (a) simply requesting participants to utter "bababa" does not change their categorical biases, but (b) prompting them to label the memoranda with 2- and 4-Labels does. Another critical observation here is that the estimated number of categories was credibly higher in the 4-Labels than 2-Labels condition. This provides further evidence that labeling had an online effect on the activation of categories that guided memory responses.

Fig. 6 presents the posterior of the category centers estimated by the free models collapsed across all participants. This figure shows that the AS and Free Labels conditions have similar number and center for the categories. The estimation of the category centers in the 2-Labels and 4-Labels conditions are similar to the one expected based on the labeling behavior imposed on those conditions. Note that the model was fitted separately to each condition and the model was blind about the manipulation of labeling, hence the estimated categories are the ones that can best account for the data in each condition.

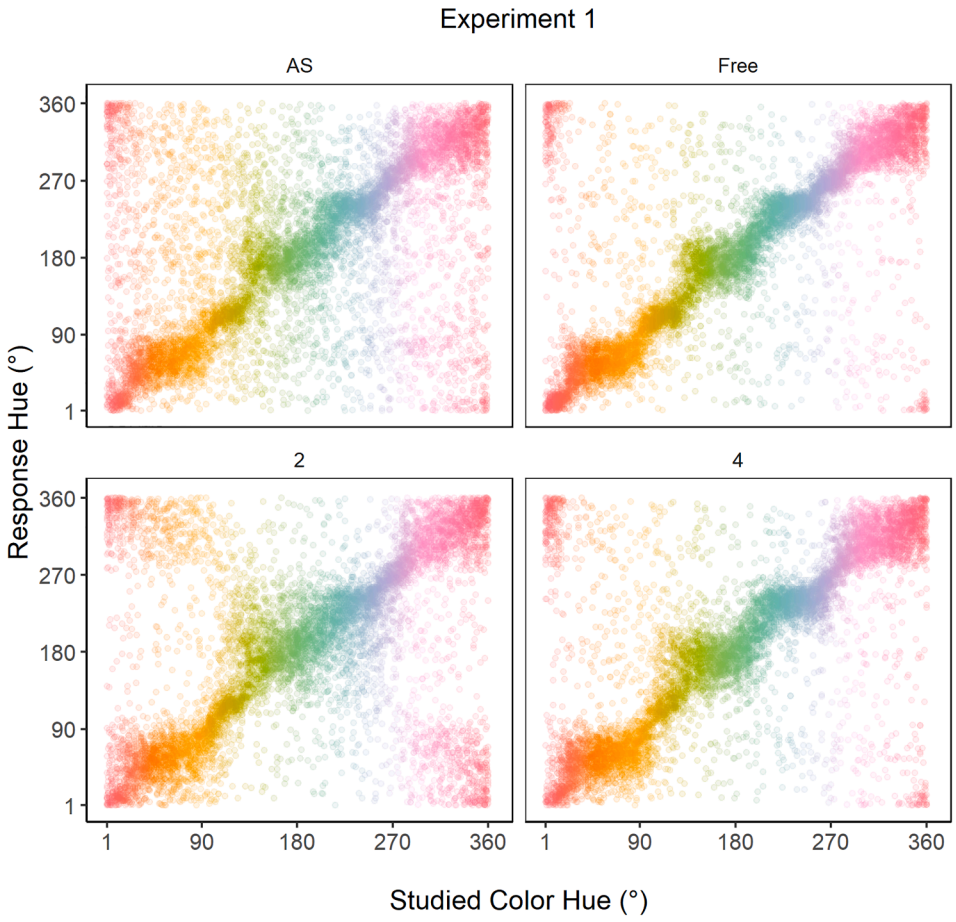

Fig. 4. Scatterplots Relating the Studied Color Hue to the Color Hue Selected as Response in Each Trial of each Experimental Condition in Experiment 1.

Probability of Recall. Fig. 5B and C show the group-level posterior parameter estimates of the fixed-categories models. Fig. 5D and E show the group-level posterior parameter estimates of the free-categories models. Fig. 5B and D present the probability that responses were informed by memory (total), and when this value is separated into the proportion of categorical and continuous representations.[2] As shown in both panels, labeling with either two, four or with free terms increased total recall probability compared to the AS condition. This was in part due to increases in categorical memory, and partially due to some increase in continuous memory. Fig. 7 presents the posterior difference in parameter estimates across the AS and each labeling condition, and also between the 2- and 4-Labels conditions in which we experimentally induced a change in label distinctiveness. When the difference between posteriors does not include 0, they are credible. As shown in Fig. 7A and B, increases in categorical memory when participants labeled the items tended to be credible, except for the contrast between AS and 2-Labels condition and the AS vs. Free condition in the fixed-categories model. Increases in continuous memory were observed for all labeling conditions in contrast to AS, except for the contrast AS-4-Labels in the Free Categories Modeling (Fig. 7C and D).

One could argue that the comparison between the AS and the 2-Labels and 4-Label conditions does not offer a proper assessment of the effect of label distinctiveness. A fairer comparison can be obtained by contrasting performance between the conditions in which label distinctiveness was manipulated experimentally, i.e., between the 2- and 4-Labels conditions. Fig. 7 also presents a direct contrast between these conditions.

When comparing the 2- and 4-Labels conditions, we can see differences in categorical and continuous memory. Categorical memory increased when the number of labels increased, but continuous memory decreased (see Fig. 7). This seems to suggest that increases in label distinctiveness reduced the storage of visual details. However, one has to take into consideration also the continuous memory imprecision parameter, the categorical imprecision, and the categorical selectivity parameters.

Imprecision. Continuous and categorical imprecision as well as category selectivity are presented in Fig. 5C and E. As shown in these panels, regardless of how the categories are set in the model (fixed or free parameter), the 2-Label condition yielded more imprecise memory compared to all other conditions, and the memory imprecision was also lower in the 4- compared to the 2-Labels condition (see Fig. 7E and F). In general, the categorical imprecision was also larger, and there was less selectivity in classifying the colors in the 2-Labels condition compared to the remaining conditions.

Discussion

Experiment 1 replicated prior work (Souza & Skóra, 2017) showing that labeling improved visual working memory performance: when labeling was prevented with an articulatory suppression procedure performance was worse compared to the Free Labels condition. Our modeling showed that this labeling benefit was due both to an increase in categorical memory – albeit in the fixed categories model this increase was not fully credible – as well as an increase in continuous memory for

[2] Note that total = categorical + continuous. For example, a total of 0.90 could be break down into 0.50 categorical and 0.40 continuous memory. The remaining 0.10 is accounted for by guessing.

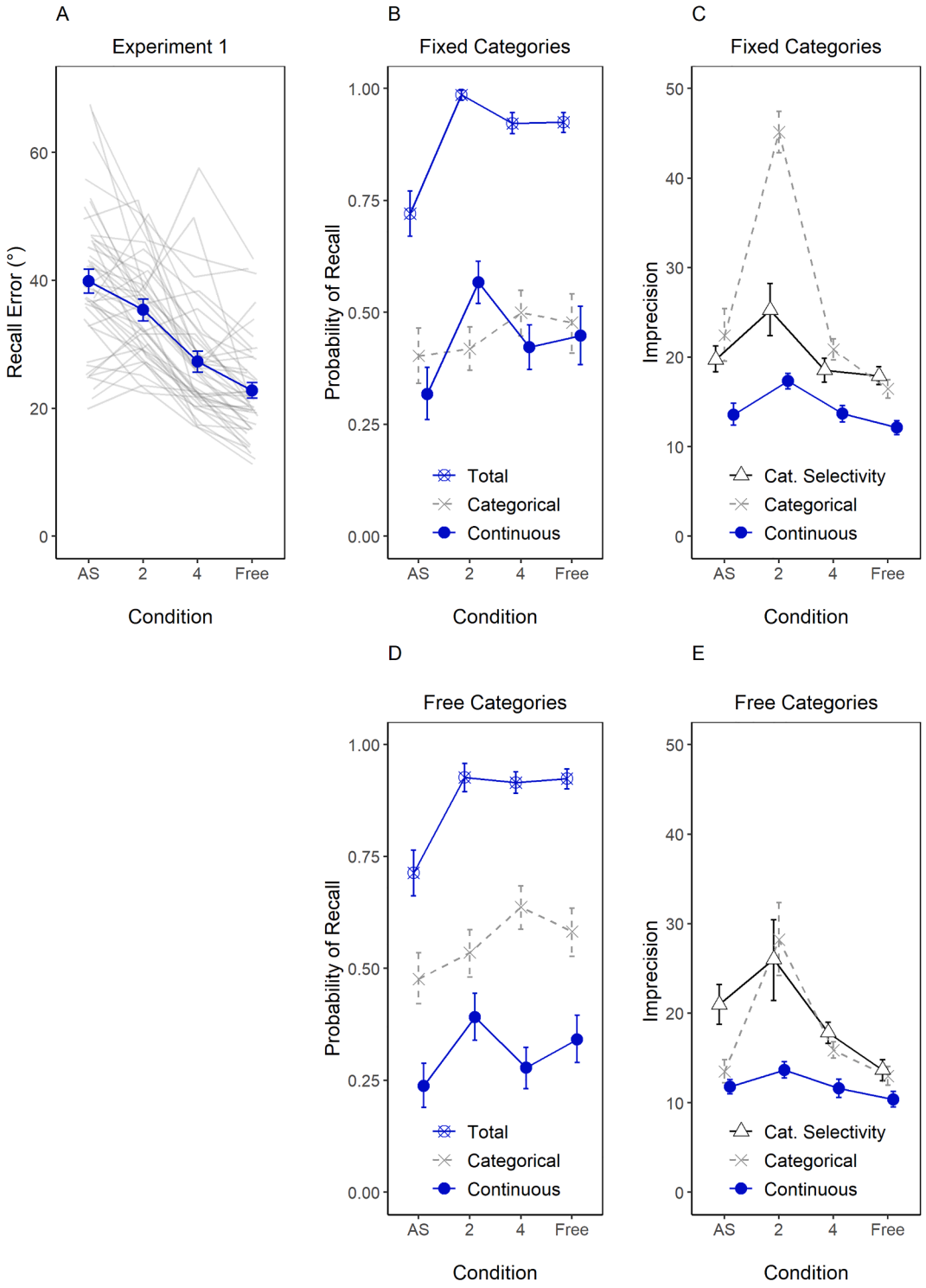

Fig. 5. Recall Error (Panel A) and Mixture Model Parameters (Panels B-E) in Experiment 1. Note. Panel A. Sample mean and error bars displaying 95% within-subjects confidence intervals are in blue; individual data in grey. Panels B and D: Mean and 95% HDI of the probability of recalling information from memory (total), and breaking this value down into categorical and continuous information. Panels C and E: Mean and 95% HDI of parameters reflecting the imprecision of continuous representations, categorical representations, and selectivity with which colors were categorized. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

the freely labeled colors. The increase in continuous memory in the Free Labels condition was reflected both in a higher probability of storing a continuous representation as well as in more precise memory. Furthermore, our modeling showed that the same categories were present when participants freely labeled the memoranda or when labeling was hindered with articulatory suppression. This is consistent with the idea that labeling is an efficient way to activate the categorical information in long-term memory (Edmiston & Lupyan, 2015; Forder & Lupyan, 2019; Lupyan & Thompson-Schill, 2012). These results are consistent with the categorical visual long-term memory hypothesis and the label-feedback hypothesis.

The novel contribution of Experiment 1 was to assess how the

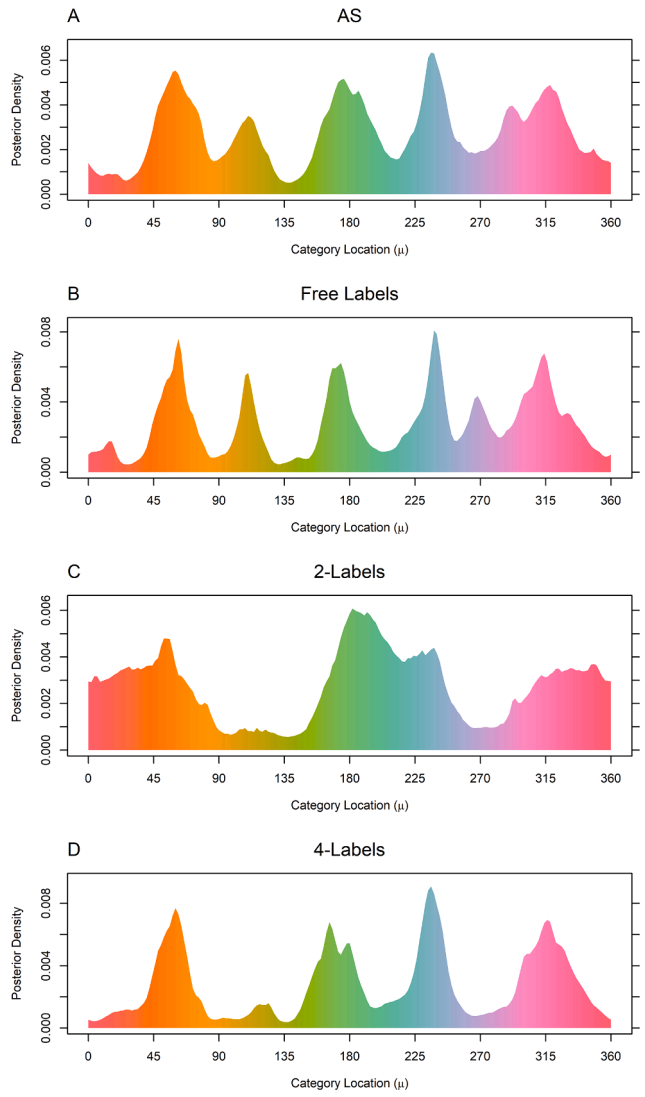

Fig. 6. Posterior Distributions of Category Centers in Each Condition Estimated by the Free Categories Models in Experiment 1.

quantity and quality of visual working memory representations change with the categorical distinctiveness of the descriptions generated in the moment to categorize the memoranda. We reasoned that if categorical information has an online effect on memory, storage of visual details would depend on how much categorical information was provided by the labels to distinguish between the different items in the memory set. To manipulate label distinctiveness, we experimentally varied the number of labels (2 or 4) used to describe the colors. Both conditions produced lower recall error compared to the suppression condition. Critically, the superiority of these conditions compared to AS was due to modulations in different parameters. On the one hand, labeling reduced guessing irrespective of the number of labels used. The chance that participants had information about the colors in mind increased dramatically ($p > 0.90$) in both labeling conditions. On the other hand, the precision of the information in memory depended on the label distinctiveness. The increase in total recall probability in the 2-Labels condition was still associated with substantial recall error, this being the case because this condition generated less precise memory than both the AS condition and the 4-Labels condition. The larger superiority of the 4-Labels condition compared to the AS and 2-Labels comes therefore from a higher chance of having information in memory and also because the more distinct labels allowed properties from the memorized hue to be better retained in mind. These findings demonstrate that labeling has a flexible and online influence on memory: it generally increases accessibility of information in memory and it will affect storage of visual details depending on the precision of the categorical information activated by the labels.

One concern often raised to our results is that the AS condition introduces task irrelevant information whereas in the other conditions all labels are task relevant. Souza and Skóra (2017) demonstrated that the AS procedure yielded similar performance as a condition in which participants generated task relevant labels (i.e., they labeled the serial input position of the items: first, second, third, and fourth). This shows that it is not the task-relevance that creates a benefit for the usage of color labels, but the categorical information they provide regarding the task-relevant feature.

Another concern that may be raised is that participants received training before entering some conditions (2-Labels and 4-Labels), but not before entering others (AS and Free Labels), and there might be carryover from one condition to the next. We only exposed participants to a training phase for the conditions that we assumed deviated from their natural tendencies to categorize the stimuli, and we counterbalanced the order of the experimental conditions such that some participants did the AS and Free Labels condition without being exposed to a different type of categorization training. Although running the risk of carryover effects, we believe the manipulation of the labeling requirements within-subjects and the findings that these conditions still had clear-cut contrasting effects on performance speaks more strongly to the online effect of labeling. These effects were obtained despite all the variation in prior experience with the experimental conditions and different labeling requirements. That said, it is worth noting that the most critical comparison regarding the effect of label distinctiveness is obtained by contrasting the 2- and 4-Labels conditions, both of which were preceded by training.

Experiment 1 allowed participants to rely on their own prior experience with color terms to generate labels that could be applied broadly or narrowly over the continuous color feature-space. Participants had very similar intuitions regarding how to label the colors, as reflected on the high agreement regarding color terms in the Free Labels condition as shown in Fig. 3C. Our results showed that labeling can have both benefits (i.e., higher chance of remembering and higher fidelity if the labels are precise) and costs (i.e., lower memory precision when labels are broad) and that this will constrain how much performance changes in comparison to a condition in which labeling is hindered. One question left unanswered in this experiment, is whether the effects of labeling depend on the long-term prior experience with the terms, or whether they mostly reflect the act of categorizing the memoranda at the study phase. The goal of Experiment 2 was to distinguish between these possibilities.

Experiment 2

In Experiment 2 participants completed a continuous shape-reproduction task (Li, Liang, Lee, & Barense, 2020). We selected this continuous space because it contains shapes that are more or less novel, and hence it was less likely that participants encountered and classified (aka. labeled) these shapes prior to the experiment. This allowed us to take a somewhat more neutral ground to experimentally manipulate the categorization of the shapes without strongly conflicting with prior knowledge of the individuals.

In Experiment 2, we experimentally built the categories used by participants in labeling the shapes. Participants were presented with German non-words, and they were trained to apply these terms to the continuous shape-space across two conditions that varied in categorical distinctiveness. In the 2-Labels condition, participants were trained in using two non-words to divide the shape space into two broad categories. In the 4-Labels condition, participants were trained in using four non-words to divide the shape space into four categories. Critically, unlike Experiment 1, the partitioning of the shape wheel was randomly

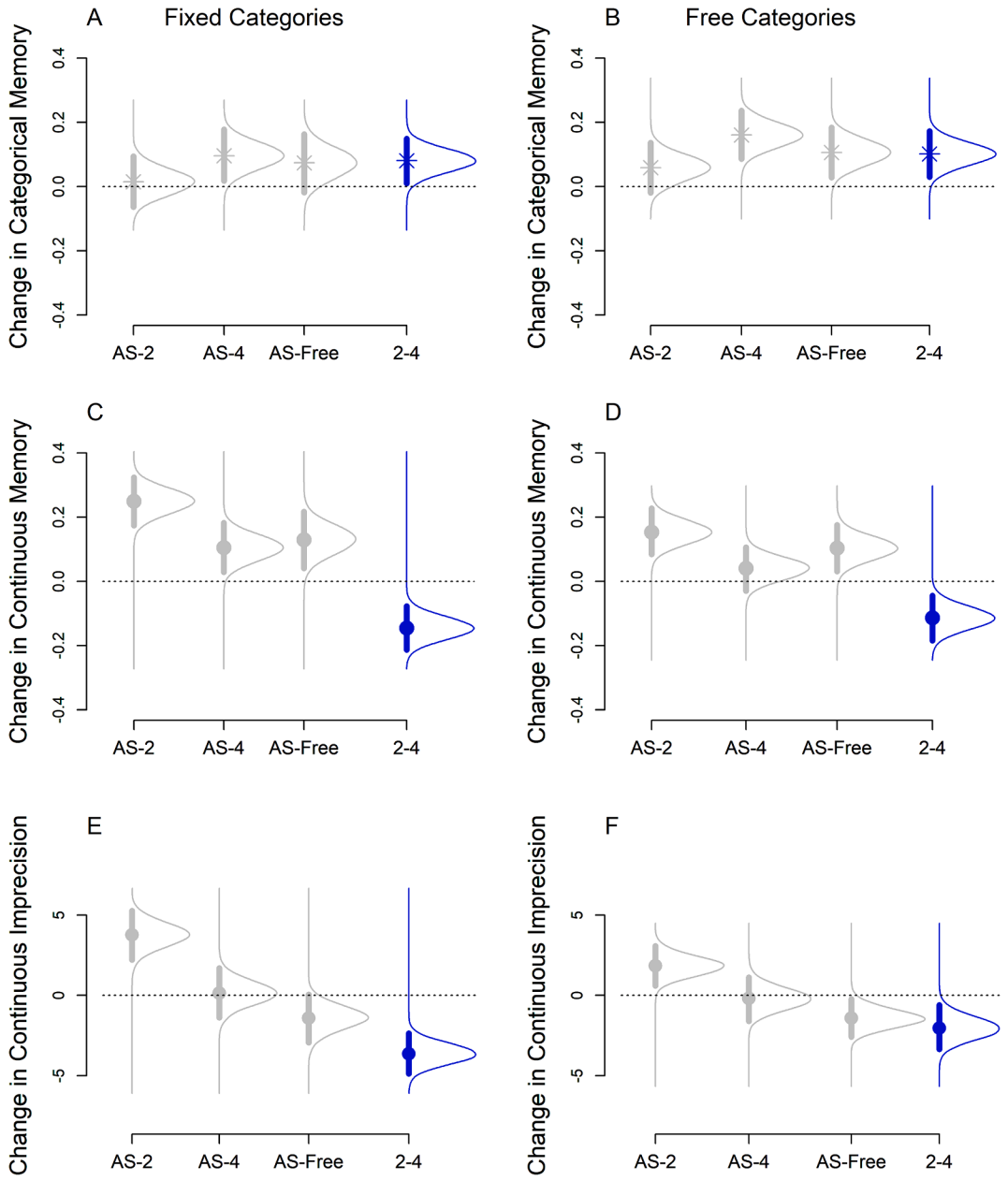

Fig. 7. Violin Plot of the Difference in Posterior Estimates Between Conditions in Experiment 1. Note. The x-axis depicts the two conditions contrasted (e.g., AS-2 = AS vs. 2). The dot represents the mean difference in estimates between the conditions and the thick line the HDI of their difference. The horizontal dotted line marks the value representing the Null hypothesis.

determined for each participant, and hence completely arbitrary. As comparisons, participants also completed the task under suppression (AS condition) and having the opportunity to freely generate labels (Free Labels condition).

If categorical distinctiveness is the relevant variable in generating the labeling effects we observed in Experiment 1, then training participants to arbitrarily categorize the stimuli using broad or more narrow categories should yield comparable effects to the ones observed in Experiment 1 although participants had no extra-laboratory experience with the labels and the categories themselves.

Methods

Participants

Thirty-two students (21 women; $M = 22$ years old) of the University of Zurich took part in two 90-min sessions in exchange of course credit or 45 CHF. All participants were German speaking individuals, and they were tested in German. One participant was excluded from the final analysis because they failed to learn the categories in the 4-Labels condition with sufficiently high accuracy (see criteria below), leaving a total N = 31. Our sample size decision was again based on the number of participants required to counterbalance our conditions (16 orders, which we replicated two times). We sought to have enough participants to detect strong evidence (BF > 10) for difference in performance between the 2- and 4-Labels condition.

Procedure

Continuous shape reproduction task. In the beginning of every trial (see Fig. 8A), four thin grey circle frames (RGB 220 220 220; 95 pixels radius) appeared evenly spaced on an imaginary circle (160 pixels radius) centered in the middle of a white screen for 500 ms. Next, one

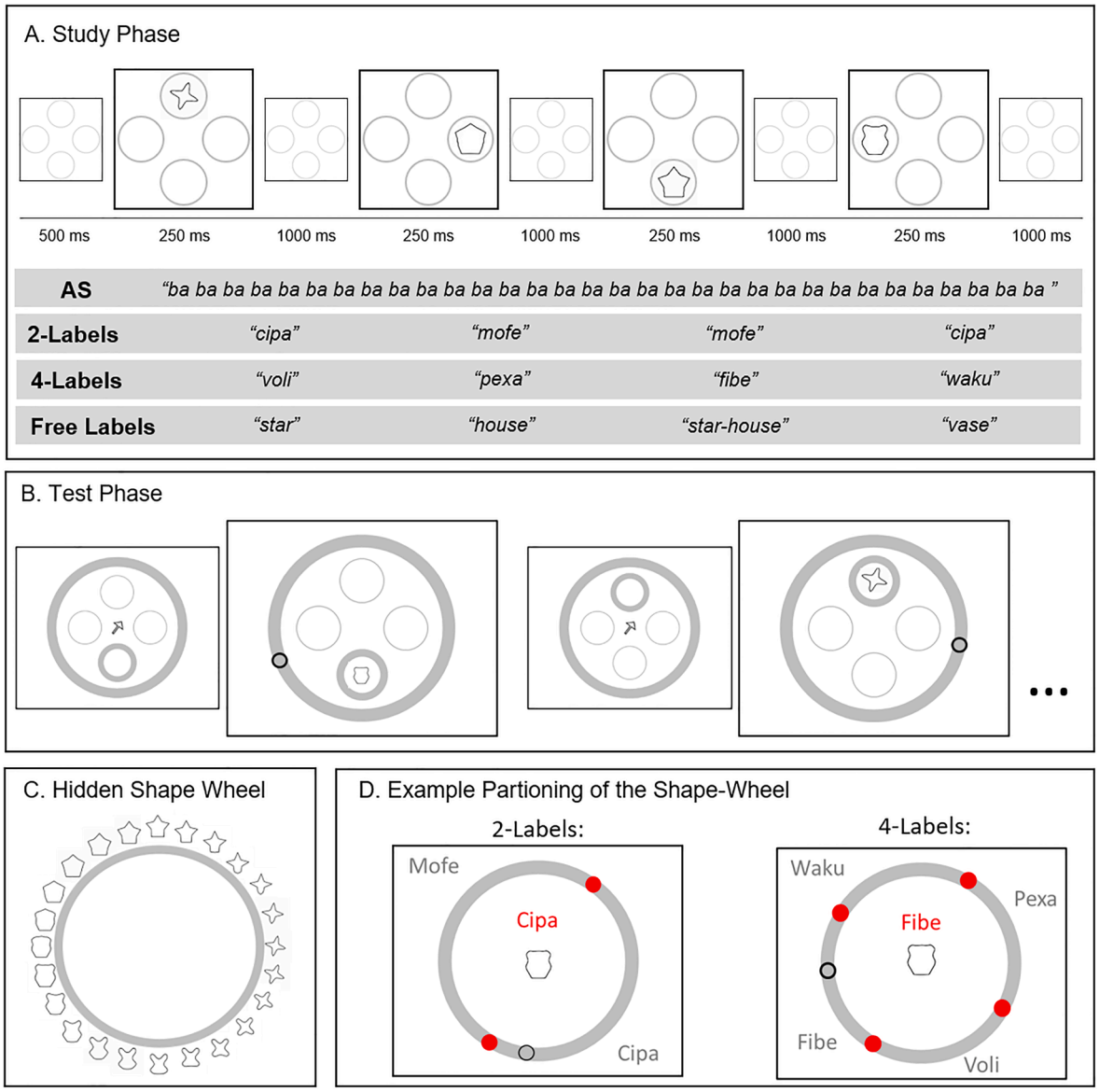

Fig. 8. Experimental Design of Experiment 2. Note. Panel A illustrates the flow of events in the study phase and the verbalization requirements in each experimental condition of Experiment 2. Panel B illustrates the test phase. The thick circle-frame indicated the item to be reproduced on the shape wheel, and the black circle indicated the current mouse position on the wheel. Panel C illustrates the hidden continuous shape wheel by displaying the shapes at evenly (15° apart) points on the wheel. Panel D illustrates the division of the shape wheel in the 2-Labels and 4-Labels conditions. The partitioning of the wheel was randomly determined for each participant in each condition.

shape was presented at one circle-location at a time for 250 ms, followed by an inter-stimulus interval of 1000 ms. Shapes were randomly sampled from 360 shapes evenly distributed along a continuous shape wheel (Li et al., 2020). The first stimulus position was randomly selected from the four locations, and the subsequent ones followed in clockwise fashion.

Next, memory for all four shapes was tested in random order. A dark-grey wheel was shown around the four grey dots, and the thicker frame of one of the circles indicated the stimulus to be reproduced (see Fig. 8B). When participants moved the mouse towards the grey wheel, one shape appeared at the tested location. Angles on the grey wheel were used to represent the shapes on a hidden-shape wheel (illustrated in simplified form on Fig. 8C). Participants moved the mouse along the wheel to adjust the shape appearing at the tested location, and they clicked with the mouse to confirm their response. The shape-wheel was randomly rotated on every trial.

As in Experiment 1, all participants were exposed to four within-subject conditions: AS condition, 2-Labels, 4-Labels, and Free Labels. Each experimental condition was completed in a different block consisting of 50 trials (see Fig. 8A). In the AS condition, participants repeated *"bababa"* aloud throughout the study phase, thereby preventing labeling. In the Free Labels condition, participants freely labeled the shapes with any term they wanted. In the 2-Labels and 4-Labels conditions, participants labeled the shapes using only two and four German non-words (see pre-training below).

All verbalizations during the memory trials were recorded, and coded offline by the experimenter. Each session consisted of only two experimental blocks. The 2-Labels and 4-Labels blocks did not occur in the same session due to time constrains related to the pre-training phase. Block order within the session and session order were counterbalanced across participants.

Label Pre-Training. Before starting the 2-Labels and 4-Labels memory block, participants underwent a three-part label categorization training phase. In the first part, participants were shown the wheel partitioned

into 2 or 4 even sections, with each section being associated with a non-word: Cipa and Mofe for the 2-Labels condition; Pexa, Voli, Fibe and Waku for the 4-Labels condition. The partitioning of the wheel was indicated with red dots (as illustrated in Fig. 8D). When participants moved the mouse around the wheel, the shape at the current mouse cursor position and its category label appeared in the screen center as illustrated in Fig. 8D. Participants were told to explore the categories and then click on a button to continue to the next part.

In the second part, participants trained categorizing the shapes. A shape was shown on the middle of the screen together with the 2 or 4 non-words reflecting the respective categories (presented in a row underneath it). Participants clicked on the label they thought applied to this shape, and received feedback regarding the correctness of their response for 1 s: the correct label turned green and, if their response was incorrect, the selected label turned red. They completed a minimum of 144 trials in this training part, which were divided into 3 blocks of 48 trials. Within each block, shapes within each category appeared evenly (to make sure training was even over categories). The phase was finished when participants achieved a minimum accuracy levels of 83% over the last block of 48 trials. In case accuracy was below that level, participants completed an additional block of 48 trials (with categories evenly balanced along these trials).

As the third and last step, participants trained verbal labeling of the items within the time-frame of the experiment (1250 ms for each item). As in Experiment 1, they were presented one stimulus for 250 ms, and had an additional 1000 ms to label it. They self-scored their label accuracy, and this training continued for a minimum of 100 trials, and until they reached 83% accuracy. Only one participant had to be excluded for failing to learn the categories (and only in the 4-Labels condition).

Results

Verbal labeling

We recorded all generated labels during the working memory trials and these were coded offline by the experimenter. To assess whether participants were correctly labeling the shapes with the trained labels, we centered the shapes in relation to the random partitioning of the

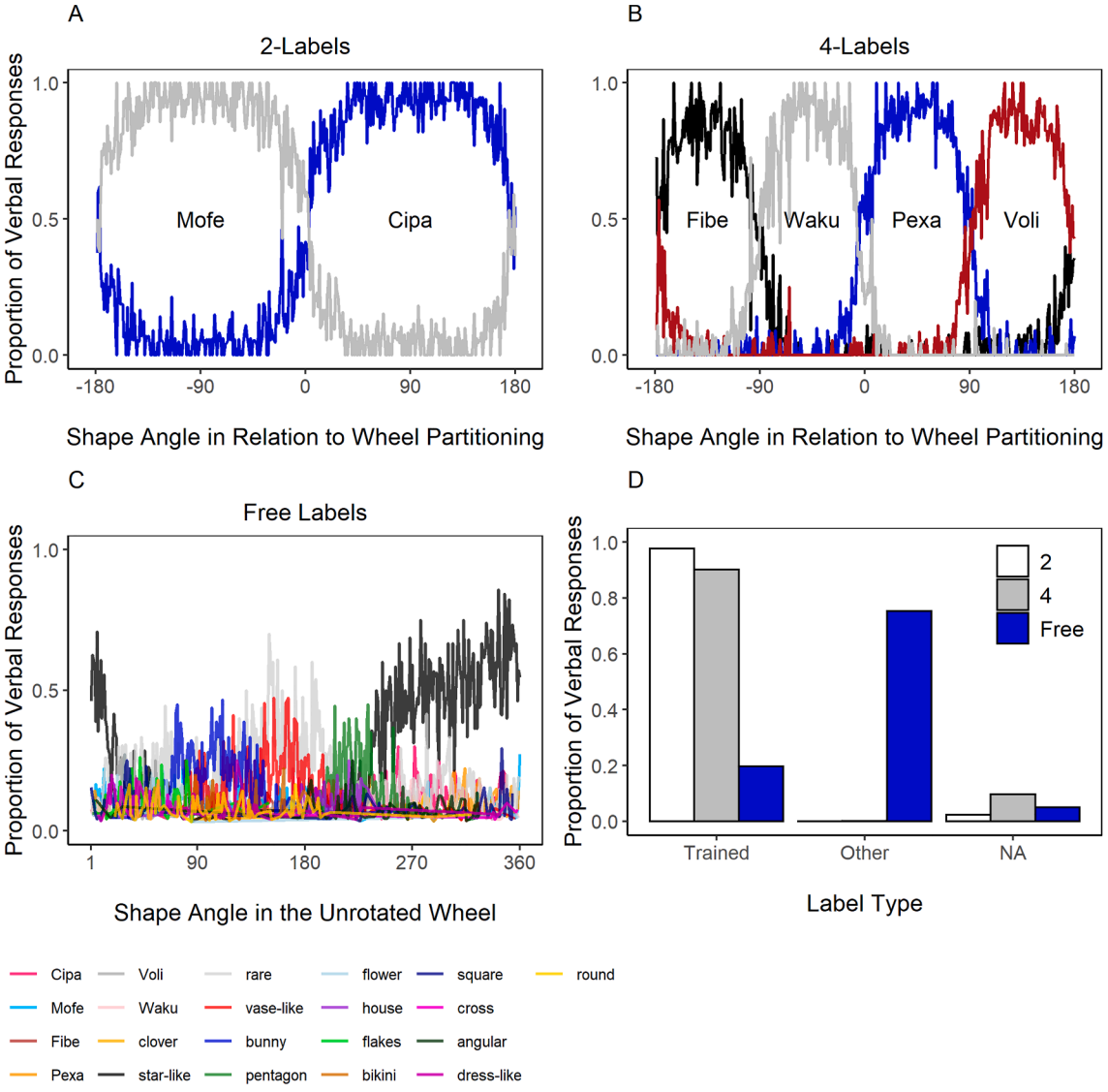

Fig. 9. Verbal Labeling Behavior Recorded During the Working Memory Trials in Experiment 2. Note. Panels A-B: Proportion of times each trained label was applied to shapes in the sections defined by the random wheel partitioning (with 0 indicating the random partitioning). Panel C: Proportion of times labels were applied to the shapes in the unrotated wheel in the Free Labels condition. Labels referring to the same quality were lumped together (e.g., small star, big star, Christmas star were classified as star-like); the first 6 labels refer to the trained non-words. Panel D: Overall usage proportion of (a) the trained labels, (b) other terms, or (c) not classified responses (NA = silence or unintelligible) across the three labeling conditions.

shape-wheel defined for each participant (which we will refer to as the 0-point) in the 2- and 4-Labels conditions. In the 2-Labels condition, the 0-point defines the separation of the two-halves of the wheel, with one half being labeled Cipa and the other Mofe. In the 4-Labels condition, the 0-point defines the start of the partitioning of the wheel into equal 90° sections which were labeled Pexa, Voli, Fibe, and Waku. Using this re-centered wheel, we computed the proportion of times the trained labels where applied to the respective shapes on the assigned sections during the working memory trials. Fig. 9A and B show one line per label, with the line tracking the proportion of times this label was applied to each shape on the wheel. As shown in Fig. 9A and B, participants correctly applied the trained labels to shapes on their respective sections of the wheel during the working memory trials. This indicates that our categorization training was effective in generating differentiated labeling behavior during the memory trials.

In the Free Labels condition, participants used 174 different labels to describe the shapes. This yielded a very low frequency of responses per shape. In order to allow for some visualization of the variety of labels used, we classified together terms that referred to similar concepts (e.g., star, dress, vase, house). We then lumped together terms that were used less frequently (less than 70 entries) into a general category of "rare". This allowed us to reduce the label-space to 21 terms.

We then plotted how these terms were applied to the shapes on the unrotated wheel (see Fig. 9C). As indicated in Fig. 9C, there was little consensus among participants on how to label the shapes: the biggest agreement was with the use of terms referring to stars in relation to shapes in angles 270 to 360 (see shape wheel on Fig. 8C, note that 0 is on the right and angles increase in clockwise fashion).

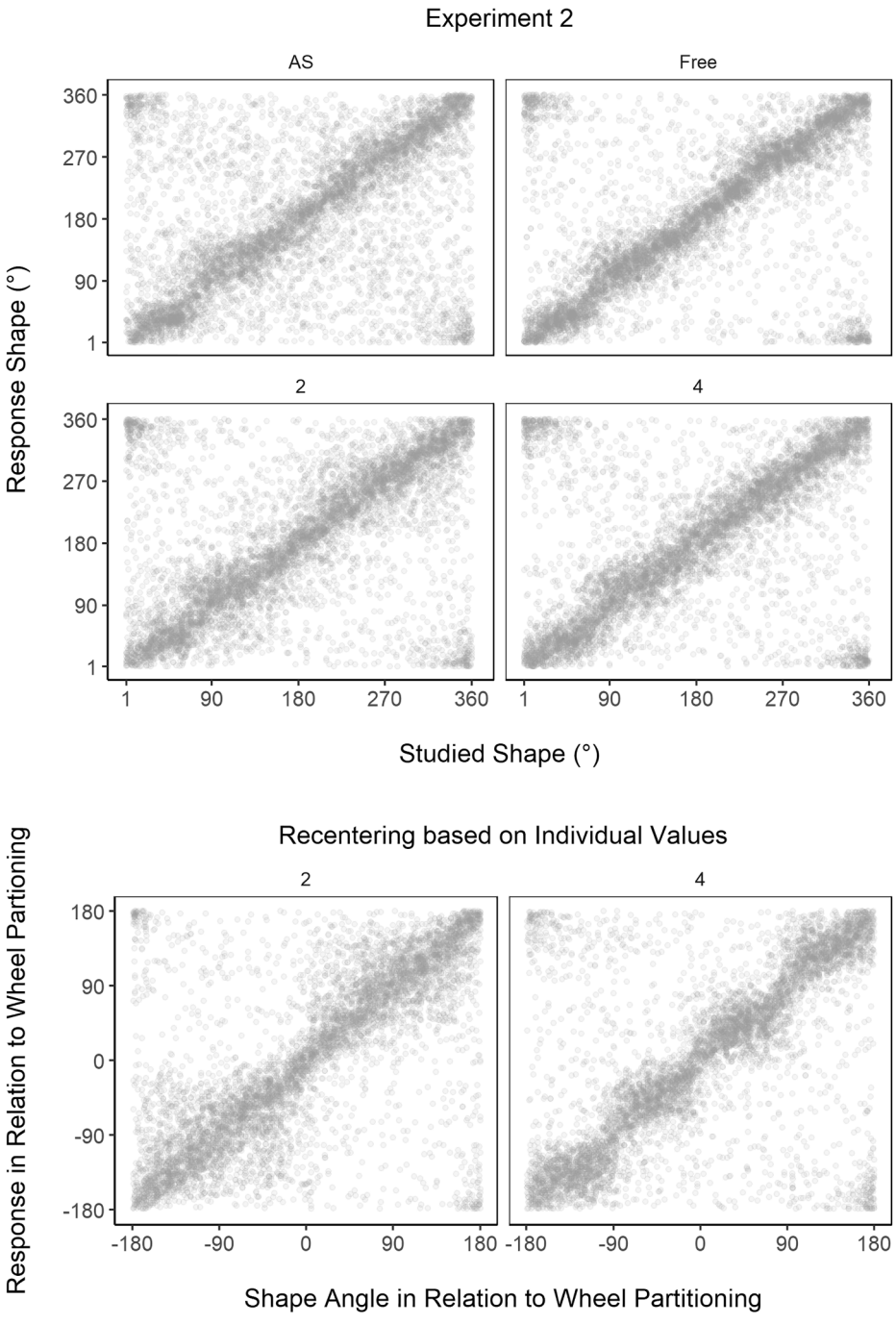

Fig. 10. Scatterplots Relating the Studied Shape to the Shape Selected as Response in each Trial of Each Experimental Condition of Experiment 2. Note. The first two rows show the data in relation to the unrotated shape-wheel. Given that wheel-sectioning was individually determined for each participant in the 2-Labels and 4-Labels conditions, the category effect is diluted in this visualization. The bottom row presents the studied shapes and response shapes in relation to the individual wheel partitioning. The labeled categories became then apparent.

Fig. 9D shows the overall proportion of times participants used the trained labels, other labels, or for which no response was recorded (silence or unintelligible sound). Participants mainly used the trained labels in the 2- and 4-Labels conditions. In the Free Labels condition, participants still used some of the trained labels (ca. 19% of the time), alongside a large range of other terms. This may be the case because 75% of the participants completed the Free Labels condition after having been already exposed to either the 2- or the 4-Labels conditions.

Model free indices of performance

Raw responses. Fig. 10 shows a scatterplot relating studied shape to response shape in Experiment 2. As anticipated, there were few categorical bands in the AS and Free Labels conditions given that the shapes do not easily map to prior learned categories. This was a desired feature in our selection of this feature space to facilitate the random manipulation of the categories in the 2- and 4-Labels conditions. Because the wheel was sectioned in a random location in the 2- and 4-Labels conditions, it is difficult to see the labeled categories in the unrotated wheel. When the studied and response values are re-centered in relation to the individual wheel sectioning (bottom row), two and four categorical bands emerge. This figure shows that training participants to arbitrarily classify the shapes into two or four categories produced a corresponding change in the way these shapes were stored in working memory.

Mean recall error. As in Experiment 1, we computed the average error in reproducing the shapes. Fig. 11A shows recall error as function of labeling condition. Replicating Experiment 1, recall error was largest when labeling was hindered with suppression (AS condition) and best when participants freely labeled the shapes (Free Labels). Performance in the 2- and 4-Labels conditions was in-between, with recall error being larger with 2- than the 4-Labels. A repeated-measures, one-way BANOVA showed overwhelming support for an effect of condition on recall error, $BF_{10} = 4.88 \times 10^9$. As in Experiment 1, we compared adjacent levels of the condition variable using *t*-tests. There was strong evidence for a reduction in recall error between the AS and 2-Labels conditions, $BF_{10} = 13.35$, and between the 2- and 4-Labels conditions, $BF_{10} = 275.74$. The evidence for a reduction in recall error between the 4- and Free Labels condition was only substantial, $BF_{10} = 3.79$.

One concern in the Free Labels condition is that in ca. 19% of the trials, one of the trained labels was used, showing some carryover between conditions. To assess whether this produced an impact on performance, we compared recall in trials in which participants used the trained labels vs. other labels. Only nine participants continued to use the trained labels in the Free Labels condition. The recall error when the trained labels were used was $M = 36.93$, 95% within-subjects CI [13.62, 60.24], and when they were not used was $M = 50.22$ [26.91, 73.54]. A Bayesian *t*-test indicated that their difference was ambiguous, $BF_{10} = 0.45$. However, if anything participants performed better when using the trained labels than otherwise. Given that we selected this shape wheel for their relative novelty in relation to the prior experience of the participants, this may indicate that about 1/3 of the participants found it hard to even come up with ad-hoc categories for the shapes during the working memory trials and found it easier to continue using the trained labels.

Mixture modeling

We submitted the data of Experiment 2 to the same mixture model as described for Experiment 1. For Experiment 2 we only fitted to the data in each condition a model in which the number and center of the

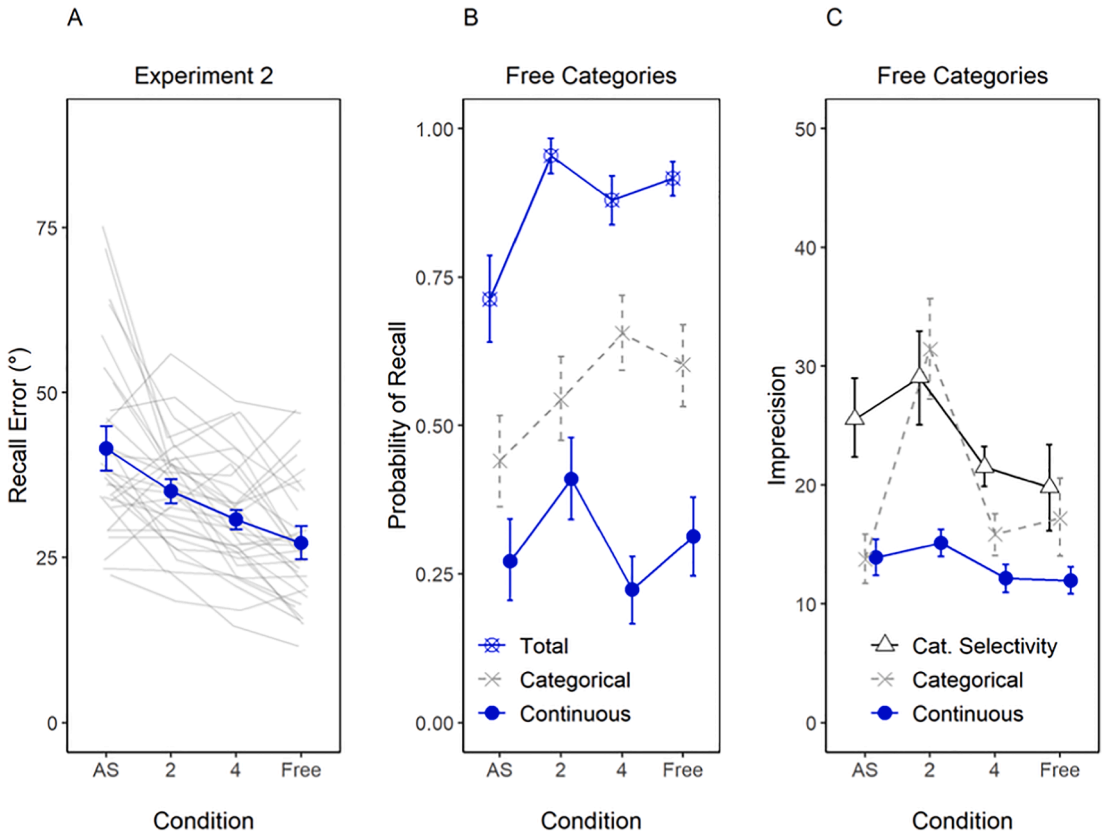

Fig. 11. Recall Error (Panel A) and Mixture Model Parameters (Panels B-C) in Experiment 2. Note. Panel A: Error bars depict 95% within-subjects confidence intervals. Grey lines depict individual participants. Panel B: Mean and 95% HDI of the group-level posterior of the probability of recalling information from memory (overall), and breaking this value down into categorical and continuous representations. Panel C: Mean and 95% HDI of the group-level posterior of the parameters reflecting the imprecision of continuous representations, categorical representations, and selectivity with which shapes were categorized.

categories were allowed to be freely estimated for each participant. Fixing the categories were not feasible here for two reasons. First, the feature space used did not involve spontaneous and stable categories which we could use to fix categories in the AS and Free Labels conditions. Second, for the 2- and 4-Labels conditions, we individually determined the partitioning of the wheel, and hence again we did not have stable categories over participants. Therefore, the best compromise was to allow the model to freely estimate these values. We fitted the model using 20,000 iterations, and discarded the first 5000 iterations as burn-in. The Supplementary Materials present a posterior predictive check of the model.

Categorical information. The number of categories estimated in each condition was the following: AS = 8.05 [7.45, 8.64]; 2-Labels = 5.52 [5.00, 6.10]; 4-Labels = 7.29 [6.84, 7.74]; and Free Labels = 7.73 [7.23, 8.29]. Only the 2-Labels condition had a credible lower number of categories than the other conditions. The remaining conditions yielded similar estimates.

Fig. 12 presents the category centers estimated by the model for a subset of participants. Averaging across all participants (as done in Experiment 1) was not informative here because participants had more idiosyncratic categories in the AS and Free Labels conditions, and their categories also differed in locations in the 2- and 4-Labels. The critical contrast here is between the 2- and 4-Labels conditions: The estimated centers show that there are fewer categories in the 2-Labels than the 4-Labels condition and the categories tend to be spaced as induced by the labeling pattern imposed on those conditions.

Probability of recall. Fig. 11B shows the group-level estimates for parameters associated with the probability of recalling information from

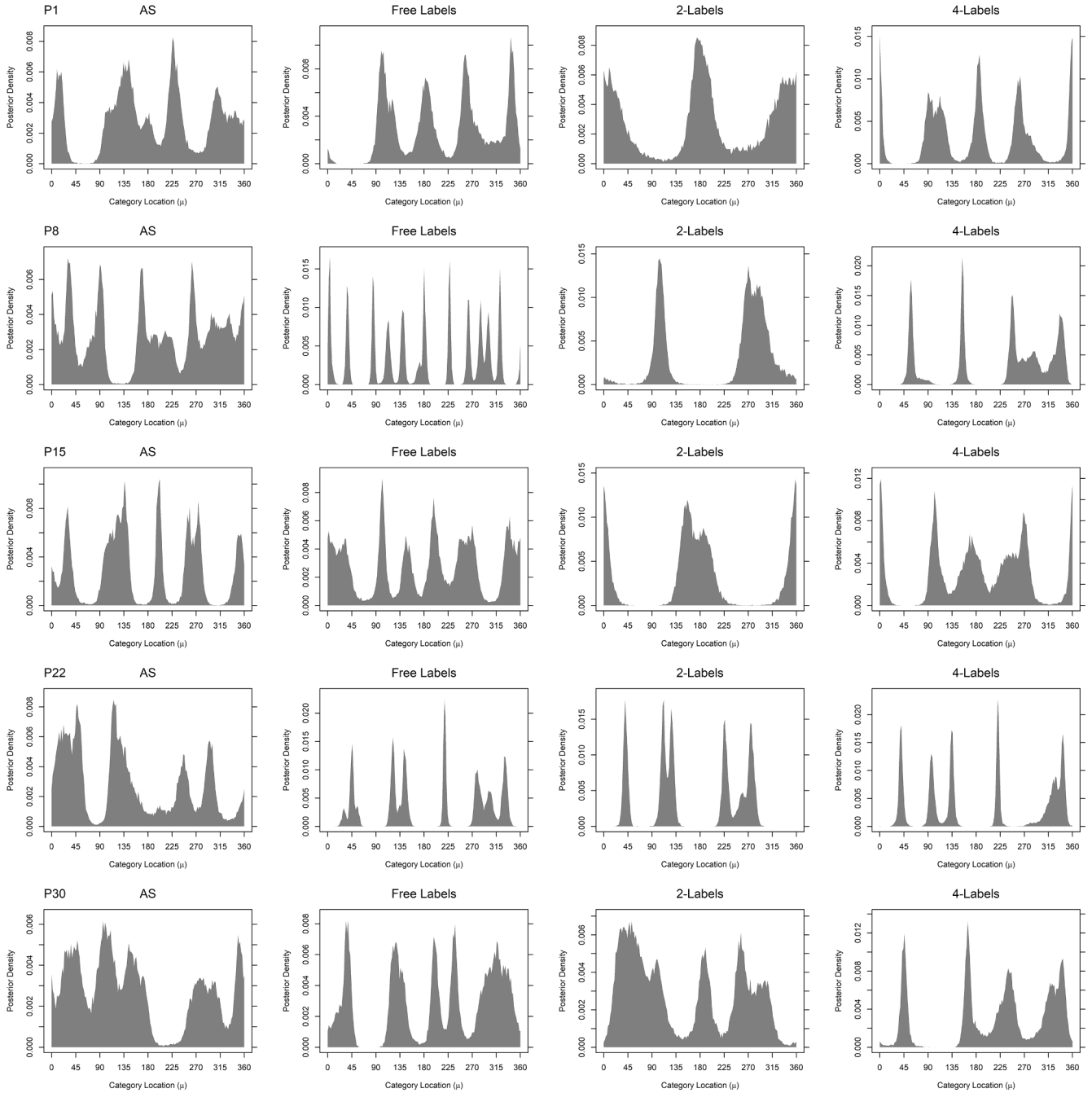

Fig. 12. Posterior Distributions of Category Centers in Each Condition for a Sample of 5 Participants (P1, P8, P15, P22, and P30) in Experiment 2.

memory. As shown in this figure, labeling with any number of terms increased the probability of recalling information from memory (total) compared to the AS condition, replicating Experiment 1. This figure also shows that this gain was mainly due to the retrieval of categorical representations. Fig. 13A shows the difference in posterior estimates for the probability of retrieving categorical information in each labeling condition against the AS condition, and when comparing the 2- and 4-Labels conditions which implement our critical manipulation of label distinctiveness. Categorical memory increased in all labeling conditions compared to AS, but this increase was not fully credible in the 2-Labels condition. The probability of categorical memory also credibly increased when contrasting the 2- and 4-Labels conditions.

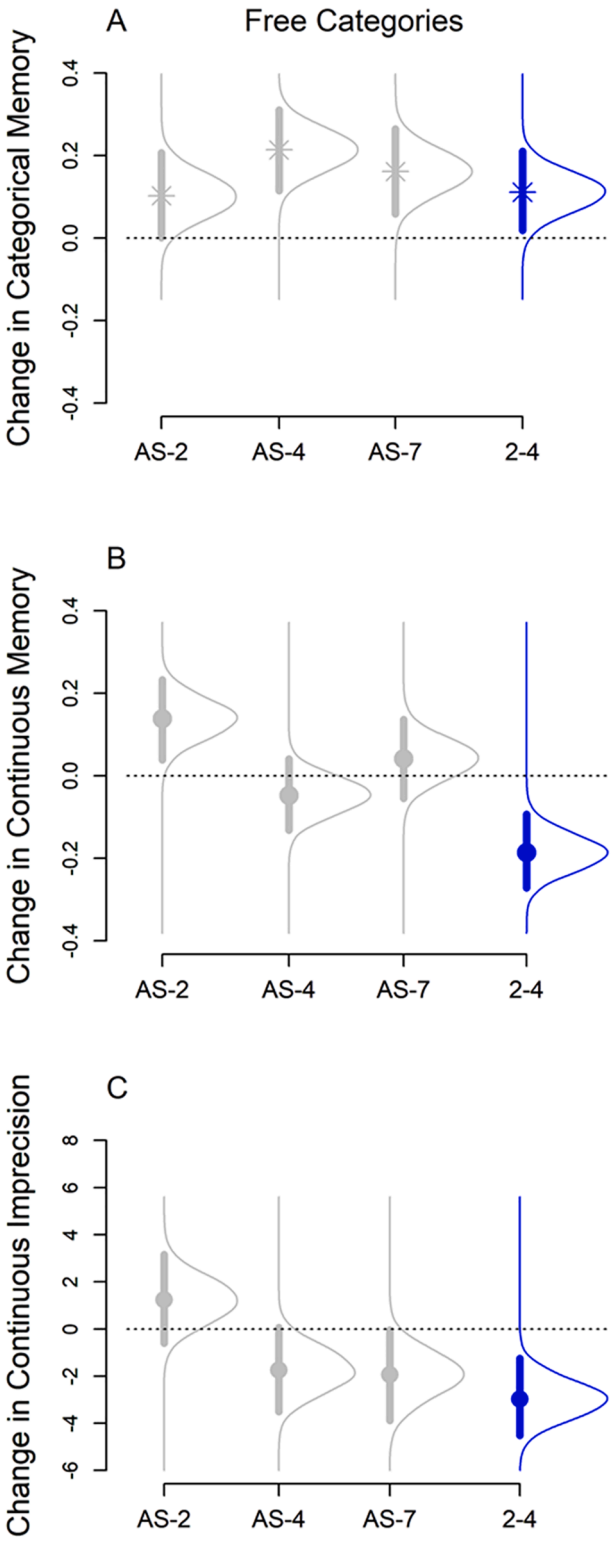

Fig. 13. Violin Plot of the Difference in Mixture Model Parameter Estimates Between Conditions in Experiment 2. Note. The dot represents the mean difference and the thick line the HDI of the difference. The horizontal dotted line marks the value representing the Null hypothesis.

Fig. 11B shows that the probability of retrieving continuous representations remained mostly unaffected by labeling, with exception of the 2-Labels condition that received a credible boost compared to AS (see also Fig. 13B). The contrast between the 2- and 4-Labels conditions showed a reduction in the probability of continuous memory in the 4-Labels condition. This is a similar pattern as observed in Experiment 1.

Imprecision. As in Experiment 1, this higher probability of recalling continuous representations in the 2-Labels condition needs to be contrasted with the reduction in memory precision this condition produced: continuous imprecision tended to increase in contrast to the AS condition (see Fig. 11C and 13C), although this increase was not fully credible in Experiment 2, and memory imprecision was credibly lower in the 4- than the 2-Labels condition, i.e., memory was more precise with 4-Labels. Likewise, categorical precision and categorical selectivity was also hindered in the 2-Labels condition compared to the other conditions (see Fig. 11C), particularly for the contrast between the 2- and 4-Labels conditions.

Discussion

In Experiment 2 we trained German-speaking individuals in categorizing shapes using German non-words. In one condition, participants learned to use two non-words to divide the continuous shape space into two arbitrary broad categories. In another condition, they learned to use four non-words to divide the same space into four arbitrary categories. Thereafter they performed a visual working memory task in which they were asked to label the shapes using these novel category labels. We compared these conditions to a suppression (AS) baseline and a free-labeling condition. Categorization of the shapes with broad (2-Labels) and more specific terms (4-Labels) had an impact on how much information was retained in visual working memory, and on the precision of memory representations. Labeling afforded the retention of more information in mind as reflected in heighted probability of retrieving information from memory when labeling was used (i.e., 2-Labels, 4-Labels, and Free Labels compared to AS). The categorical distinctiveness of the labels, however, impacted memory precision: usage of broad labels (2-Labels) hindered precision compared both to AS (although this effect was not credible in Experiment 2) and to the usage of more distinct labels (4-Labels condition). These effects replicate the ones obtained in Experiment 1 in which the labels were self-selected by participants and reflected their prior history with color terms. Here we demonstrated that this effect reflects ongoing categorization of the visual input: novel categories learned in the course of a brief training session can also serve as a prior for storage of information in visual working memory. In sum, Experiment 2 shows that labeling affects visual working memory not only when well-established categories in long-term memory are activated, but also when newly learned categories are activated by the labels. This shows that labeling has an online and task-dependent effect on visual working memory.

One apparent limitation of Experiment 2 is that participants had very idiosyncratic labeling for the shapes in the Free Labels condition and their labeling was not highly consistent. This probably limited the benefit they obtained in the Free Labels condition. This may explain why there was just a small difference between the recall error in the 4-Labels and Free Labels condition. Some participants even continued to label the stimuli with the trained labels, showing carryover effects from the previous condition they were exposed to. We don't believe this is critical

for our research question though. Here our main goal was to manipulate the categorical distinctiveness for a set of relative novel stimuli for which the categories themselves were arbitrarily constructed during the experiment. We implemented this manipulation across the 2- and 4-Labels conditions. We succeeded in replicating the results of Experiment 1 using arbitrary categories: probability of recall increased in both labeling conditions, but fine-grained memory of the shapes was reduced in the 2-Labels compared to 4-Labels conditions.

General discussion

Previous research has shown that labeling visual inputs allows us to better remember them over short intervals (Forsberg et al., 2020; Souza & Skóra, 2017). Here we demonstrated for the first time that the effect of labeling upon visual working memory is directly related to the categorical distinctiveness of the labels generated. Labels that are broadly applied to categorize the visual stimuli improve the accessibility of information in memory (increasing probability of recall) while at the same time hindering memory for visual details, whereas labels that more narrowly distinguish between items improve both the accessibility of memory representations and the retention of visual details.

Hypotheses of the labeling effect

The hypotheses of the labeling effect make different predictions to performance in our task. The verbal recording hypothesis predicts that labeling should reduce the storage of continuous information, because the label replaces the continuous information in mind. The memory distortion hypothesis predicts that labeling will bias memory towards the prototype, thereby reducing memory precision. The dual-trace hypothesis predicts that labeling will only add categorical information with no impact on the continuous information stored. The distinctiveness hypothesis predicts that any type of label should increase probability of recall as long as they provide an additional retrieval cue to the memory traces. Finally, the categorical visual long-term memory hypothesis predicts that labeling will activate categorical information in long-term memory which will serve as a categorical prior to reduce uncertainty in the incoming perceptual information (in line with the label feedback hypothesis). This reduction in uncertainty sharpens the perception and consequent storage of the visual details in working memory.

Our results show that labeling continuously varying colors and shapes can increase memory precision ruling out hypotheses that do not consider the possibility of a gain in this parameter – namely the recoding, distortion and dual-trace hypotheses. This replicates and extends to other visual features prior work with labeling of colors (Forsberg et al., 2020; Souza & Skóra, 2017). The pure distinctiveness account was also ruled out by Souza and Skóra (2017): simply using different terms to refer to the memoranda did not improve performance when they lacked categorical information. Altogether, the improvement of memory precision by labeling is consistent with the categorical visual long-term memory hypothesis.

The prediction that labels activate categorical information leads to the question of whether this activation is flexible and task-dependent. The same stimuli can be categorized in different ways by the same person. If labels are biasing storage online, then their effect will depend on which categories are activated by the labels. To the best of our knowledge, we are the first to experimentally manipulate the categorical distinctiveness of the labels and to assess its impact on visual working memory.

Categorical distinctiveness

We trained participants to categorize colors and shapes using either two or four labels. This allowed us to show that the improvement in visual working memory (i.e., the reduction in recall error) was proportional to the number of labels used. Two labels produced a smaller benefit than four labels in comparison to a verbal suppression condition. The smaller benefits of two compared to four labels is related to two opposing effects created by labeling: labels increase probability of recall indicating that labeled information is better consolidated or maintained in working memory; however, the degree in which the labels sharpen perception of the incoming information depends on how much the label highlights features that are characteristic of the category in comparison to other categories. Broad categories have fewer distinctive features than narrow categories, and hence they do not create a sufficiently distinct context to encode the precise feature of the studied items.

The impact of categorical distinctiveness uncovered here is line with the one predicted based on findings obtained in studies on episodic visual long-term memory (e.g., Richler et al., 2013). Our study advances this literature by showing that categorical distinctiveness effects can be observed by varying the types of labels (2 vs. 4 labels) applied to the same set of visual stimuli, and by comparing it to a baseline in which labeling is prevented by articulatory suppression.

In episodic visual long-term memory labeling has been associated with either costs or, at best, no changes in performance (Kelly & Heit, 2017; Lupyan, 2008; Richler et al., 2013). In contrast, our study points to a labeling benefit (see also Forsberg et al., 2020; Souza & Skóra, 2017). The differences in costs *versus* benefits might be related to the selection of baseline: long-term memory studies have typically compared verbal labeling to preference rating. However, preference rating yields better performance compared to other types of encoding instructions (Blanco & Gurenckis, 2013). By varying the labeling behavior applied to the same set of memoranda, we could provide a clearer measure of the online effect of labeling on visual working memory. Our findings showed that in contrast to a condition in which labeling is hindered, we can measure benefits even of broad labels. These benefits are smaller, however, than when more distinct labels are used.

Labeling effect: memory quantity vs. quality

Our modeling of the data provided information about the changes induced by labeling regarding the quantity and quality of visual working memory representations. Overall, labeling the task-relevant feature benefits visual working memory irrespective of the type of label (even if very broad) by increasing storage probability compared to when labeling is hindered. The benefits of very broad labels are smaller, however, because they diffusely highlight the properties of the category and hence they do not strongly sharpen perception of the incoming information. In contrast, distinctive labels have narrower boundaries and they can be used to highlight the similarity of the current item with the category prototype and its dissimilarity to other categories as suggested by the label-feedback hypothesis (Lupyan, 2012b, 2012a). This in itself may permit the storage of more fine-grained representations or more stable fine-grained representations – which are reflected in reductions in the memory imprecision parameter of our model.

These results resonate with the ones of Richler et al. (2013) in which the usage of low-distinctiveness labels (two categories) yielded worse performance than preference rating, whereas the usage of unique, high-distinctive labels (unique categories) yielded comparable performance to preference rating. Our results point to the exact source of this effect: categorical distinctiveness changes the fidelity of the memory representation. When one needs to compare several exemplars from the same category, memory fidelity becomes critical to distinguish between targets and lures. With fewer terms and several exemplars, memory imprecision will increase as a function of labeling. With several terms and few exemplars, benefits will tend to accrue.

Notwithstanding the widespread hypothesis that labels only contribute categorical knowledge (Cibelli et al., 2016; Hardman et al., 2017), our study shows that labels also affect the storage of continuous

memory representations in memory. Fine-grained visual representations seem particularly fragile, and the categorical information activated by the labels may help in protecting it from interference produced by the other items or by decreasing the load on the limited capacity of visual working memory. This protection however depends on the distinctiveness of the labels used. This goes one way in explaining why sometimes labeling could appear to produce a performance cost. Whether there will be benefits or costs will depend on the match between the types of labels used to categorize the stimuli and the subsequent demands on memory precision in the memory test. If the labels have low categorical distinctiveness and the memory test requires fine-grained discriminations, costs will likely follow. Otherwise, labels will tend to increase performance as indicated by the increased probability of retention of information in memory and the increase in memory fidelity.

Are categories and labels the same Thing?

One may wonder whether labeling and categorization are the same thing. Here we have used these terms almost interchangeably, but this is not meant to reflect that labeling is a precondition for categorization. As we have pointed out in the introduction, categorical effects are observed even in the absence of overt labeling (Crawford et al., 2000; Huttenlocher et al., 1991) and can occur as fast as 150–200 ms after stimulus onset as revealed by EEG differentiation between stimuli associated with the same vs. different labels (Thierry et al., 2009). Categorical responding appears in the delayed estimation task even under articulatory suppression as demonstrated here and in prior work (Forsberg et al., 2020; Souza & Skóra, 2017). Labels just accentuate the categorical responding in this task, a finding that has been replicated in many other perceptual tasks as well (Boutonnet & Lupyan, 2015; Edmiston & Lupyan, 2015; Forder & Lupyan, 2019; Lupyan & Thompson-Schill, 2012). In support of this notion, our modeling showed similar categorical effects under suppression and free labeling indicating that categorical information is also activated by the stimulus itself. The stimulus probably prompts the labeling, with the labeled category then more strongly activating the category in visual long-term memory which further sharpens memory storage – reflecting a cycle of interdependency.

Overall, our findings point to a strong role of in-the-moment verbal labeling for storage of visual information. Much earlier work has been concerned with comparing performance for hard-to-label *versus* easy-to-label stimuli under the assumption that the activation of labels occurs automatically upon presentation of the stimulus (Brandimonte, Hitch, & Bishop, 1992; Gilbert, Regier, Kay, & Ivry, 2006; Thierry, Athanasopoulos, Wiggett, Dering, & Kuipers, 2009). Our work shows that labeling effects are not solely automatic – categorical information is more strongly activated when labels are generated online (Lupyan, 2012b, 2012a). Our results show that it is not only the existence of a label to the visual input that matters, but whether and which type of label (more or less categorically distinct) is applied at the incoming visual information. This resonates with the idea that categorical effects are not fixed. Humans can entertain different categories for the same set of stimuli depending on the purpose of the current task. This flexibility, however, also means that what will be remembered from this episode will vary depending on the categorization.

Furthermore, our results show that the labeling effect does not need to come from a long history of culturally-mediated learning of the categories and the labels. In Experiment 1, participants could rely on their prior history with color terms to classify the colors in the Free Labels condition, and our sectioning of the color wheel in the 2-Labels and 4-Labels conditions was not arbitrary but meant to maximize consistency within the sections to facilitate the labeling and categorization of the colors. In Experiment 2, we removed the grounding on prior experience. The categories were learned over the course of a short experimental training phase that took ca. 20–30 min and the category boundaries were randomly determined for each individual. The effect of categorical distinctiveness was remarkably similar between experiments.

Do the categories come from using language to refer to visual input? Category learning is facilitated by labels, a finding that has been demonstrated with adults (Lupyan et al., 2007), children (Sloutsky & Fisher, 2012), and babies (Althaus & Mareschal, 2014). This does not mean that categorical learning cannot happen in the absence of labeling. Categorical learning probably requires discriminative training to respond differently to different categories and similarly to elements from the same category. Teaching someone to label the categories adds one property to discriminate between category elements and increases the experience in differentiating them.

Labeling and verbal rehearsal

One may wonder whether the labeling effect we observed is related to verbal rehearsal of the labels. Verbal labeling and verbal rehearsal are not the same process: labeling is related to the recoding of visual information into a verbal representation; conversely, verbal rehearsal is the overt or covert repetition of verbal information to oneself when this information has to be maintained in mind (Baddeley, 1986). Although different, once participants labeled the visual items, the verbal labels would be available for rehearsal. When participants used fewer labels to categorize the stimuli, rehearsal could be less efficient since it was more likely that multiple items received the same label, being then rehearsed together. This could in turn lead to a higher chance of confusing items labeled with the same term with each other, generating the so-called swap errors. Swap errors have been observed to account for some proportion of the errors in visual working memory tasks (Bays et al., 2009; Oberauer et al., 2017; Souza et al., 2014; van den Berg et al., 2014).

We think this explanation is unlikely for two reasons. First, our model estimated high chances of correct recall in all labeling conditions compared to suppression. This heightened probability of recall indicates that participants were not confusing items with each other, since swap errors would be accounted for in our model as guessing. Second, we plotted the distributions of responses in relation to the target feature value (i.e., the one that should be recalled now) and in relation to the non-target values (i.e., the feature of the other memory items that are not being tested now). These plots should reveal whether there is substantial concentration of responses around the non-target features which would be consistent with large proportions of swap errors. These distributions are presented in Fig. 14.

Fig. 14 shows that there was little evidence for a contribution of non-target recalls in all of the conditions (see small bump in the red distributions). Labeling with fewer terms (2-Labels) did not increase this bump compared to suppression. Of course, since performance improved across the labeling conditions, there was a smaller chance of errors in these conditions, so there was a very small reduction of the non-target bump in the 4-Labels and Free Labels conditions compared to the other conditions.

Furthermore, it is so far unclear whether verbal rehearsal would be an important contributor to performance on visual working memory tasks. The role of verbal rehearsal has been questioned even for the maintenance of verbal information with the evidence bearing this contribution being weak at best (Lewandowsky & Oberauer, 2015; Oberauer, 2019). Souza and Oberauer (2018, 2020) have experimentally manipulated rehearsal of verbal lists in two types of verbal working memory tasks (i.e., simple span and complex span) and observed that increasing the amount and length of rehearsals did not lead to any performance improvement. This is inconsistent with a causal role of verbal rehearsal in verbal tasks. Our current experiments do not rule out a role of verbal rehearsal in visual working memory, but in light of the evidence contrary to it being a helpful strategy in verbal tasks, it seems unlikely that rehearsal contributes to the labeling benefit in visual working memory.

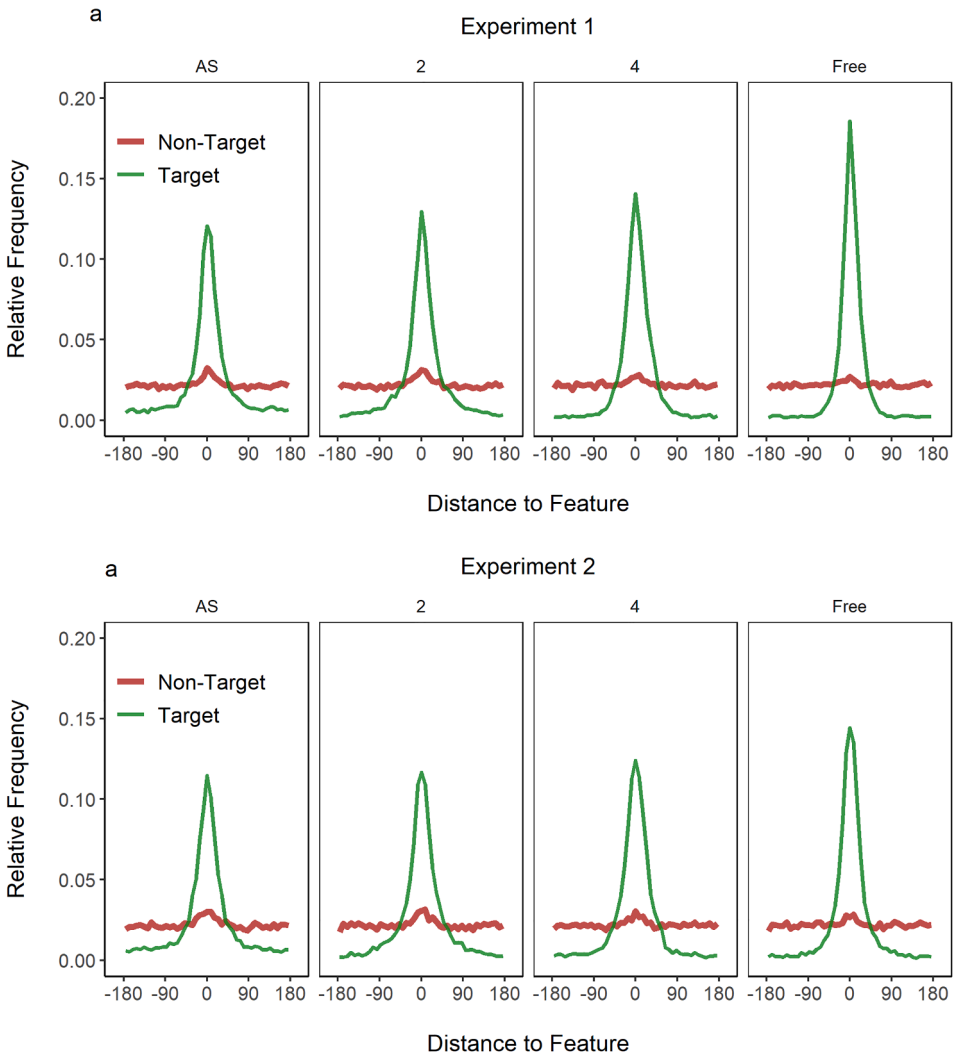

Fig. 14. Distribution of the Distance between the Response and the Target (in green) and Non-Target (in red) Features in Experiments 1 and 2. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

High-order information in working memory

Recently, studies have started to incorporate high-order information into models of visual working memory. These studies showed that information stored in working memory is context-dependent: memory for one element depends on the surrounding elements, and general statistics describing the assemble of memory items can improve memory retention (Bates, Lerch, Sims, & Jacobs, 2019; Brady & Alvarez, 2011, 2015; Brady & Tenenbaum, 2013). Furthermore, prior knowledge of co-occurrences between features seem to afford data-compression or chunking (Brady, Konkle, & Alvarez, 2009; Gobet et al., 2001; Huang & Awh, 2018; Nassar, Helmers, & Frank, 2018), which also reduces the load on working memory. Similarly, it is conceivable that labeling and categorization may also allow participants to use hierarchical representations and contextual information to increase data-compression thereby relieving capacity limitations in visual working memory. This is in line with our prior working demonstrating that labeling had a substantially larger effect on visual working memory when the memory load increased from one to two, and then to four items (Souza & Skóra, 2017). When participants had to retain one single element, labeling was inconsequential, and with two-items the effects were also small. The benefits were only substantial when four items were retained in mind. Storage of multiple items places a strong demand on the limited capacity of visual working memory, and it is under these conditions that labeling is most beneficial.

In sum, our results are relevant for theories regarding the interplay of visual working memory and conceptual long-term memory (Brady, Konkle, & Alvarez, 2011). They point to ways in which activation of conceptual long-term memory via labeling changes online storage of information in visual working memory providing evidence that the interplay between these two systems can be under strategic control of the individual.

Conclusion

The richness of a picture may be worth a thousand words, but the retention of this complex visual information in mind is constrained by capacity limitations. Our study shows that investing more words in describing a picture can lead to a big payoff: visual representations become more accessible and they retain more fine-grained details.

CRediT authorship contribution statement

Conceptualization: ASS conceptualized Experiment 1 together with MM. ASS conceptualized Experiment 2 together with CO. Data curation: ASS worked on the data curation. Formal analysis: CO and MM worked on the data analysis of Experiment 1, and ASS analyzed the data of Experiment 2. Funding acquisition: ASS obtained funding to support this

research. Investigation: MM collected data of Experiment 1, and research assistants under supervision of ASS (namely, Gian-Luca Gubler and Moritz Truninger) collected data of Experiment 2. Methodology, Project administration, Resources, Supervision, and Visualization were done by ASS. Writing - original draft: ASS. Writing - review & editing: MM and CO.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This research was supported by a grant from the Swiss National Science Foundation (grant n° 169302) to A. S. Souza.

Appendix A. Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jml.2021.104242.

References

Alogna, V. K., Attaya, M. K., Aucoin, P., Bahník, S., Birch, S., Birt, A. R., … Zwaan, R. A. (2014). Registered replication report Schooler and Engstler-Schooler (1990). *Perspectives on Psychological Science, 9*(5), 556–578. https://doi.org/10.1177/1745691614545653.

Althaus, N., & Mareschal, D. (2014). Labels direct infants' attention to commonalities during novel category learning. *PLOS ONE, 9*(7), Article e99670. https://doi.org/10.1371/journal.pone.0099670.

Athanasopoulos, P., Damjanovic, L., Krajciova, A., & Sasaki, M. (2011). Representation of colour concepts in bilingual cognition: The case of Japanese blues. *Bilingualism: Language and Cognition, 14*(Special Issue 01), 9–17. https://doi.org/10.1017/S1366728909990046.

Baddeley, A. (1986). *Working memory*. Clarendon Press/Oxford University Press.

Bae, G. Y., Olkkonen, M., Allred, S. R., & Flombaum, J. I. (2015). Why some colors appear more memorable than others: A model combining categories and particulars in color working memory. *Journal of Experimental Psychology: General, 144*(4), 744–763. https://doi.org/10.1037/xge0000076.

Bae, G. Y., Olkkonen, M., Allred, S. R., Wilson, C., & Flombaum, J. I. (2014). Stimulus-specific variability in color working memory with delayed estimation. *Journal of Vision, 14*(4), 7. https://doi.org/10.1167/14.4.7.

Bartlett, J. C., Till, R. E., & Fields, W. C. (1980). Effects of Label Distinctiveness and Label Testing on Recognition of Complex Pictures. The American Journal of Psychology, 93(3), 505–527. JSTOR. https://doi.org/10.2307/1422727.

Bates, C. J., Lerch, R. A., Sims, C. R., & Jacobs, R. A. (2019). Adaptive allocation of human visual working memory capacity during statistical and categorical learning, 11–11 *Journal of Vision, 19*(2). https://doi.org/10.1167/19.2.11.

Bays, P. M., Catalao, R. F. G., & Husain, M. (2009). The precision of visual working memory is set by allocation of a shared resource. *Journal of Vision, 9*(10), 7. https://doi.org/10.1167/9.10.7.

Blanco, N., & Gureckis, T. (2013). Does category labeling lead to forgetting? *Cognitive Processing, 14*(1), 73–79. https://doi.org/10.1007/s10339-012-0530-4.

Boutonnet, B., & Lupyan, G. (2015). Words jump-start vision: A label advantage in object recognition. *The Journal of Cognitive Neuroscience, 35*(25), 9329–9335.

Bower, G. H., Karlin, M. B., & Dueck, A. (1975). Comprehension and memory for pictures. *Memory & Cognition, 3*(2), 216–220. https://doi.org/10.3758/BF03212900.

Boynton, R. M., Fargo, L., Olson, C. X., & Smallman, H. S. (1989). Category effects in color memory. *Color Research & Application, 14*(5), 229–234. https://doi.org/10.1002/col.5080140505.

Brady, T. F., & Alvarez, G. A. (2011). Hierarchical encoding in visual working memory ensemble statistics bias memory for individual items. *Psychological Science, 22*(3), 384–392. https://doi.org/10.1177/0956797610397956.

Brady, T. F., & Alvarez, G. A. (2015). Contextual effects in visual working memory reveal hierarchically structured memory representations. *Journal of Vision, 15*(15), 6. https://doi.org/10.1167/15.15.6.

Brady, T. F., Konkle, T., & Alvarez, G. A. (2009). Compression in visual working memory: Using statistical regularities to form more efficient memory representations. *Journal of Experimental Psychology: General, 138*(4), 487–502. https://doi.org/10.1037/a0016797.

Brady, T. F., & Tenenbaum, J. B. (2013). A probabilistic model of visual working memory: Incorporating higher order regularities into working memory capacity estimates. *Psychological Review, 120*(1), 85–109. https://doi.org/10.1037/a0030779.

Brainard, D. H. (1997). The psychophysics toolbox. *Spatial Vision, 10*(4), 433–436. https://doi.org/10.1163/156856897X00357.

Brandimonte, M. A., Hitch, G. J., & Bishop, D. V. M. (1992). Verbal recoding of visual stimuli impairs mental image transformations. *Memory & Cognition, 20*(4), 449–455. https://doi.org/10.3758/BF03210929.

Brandimonte, M. A., Schooler, J. W., & Gabbino, P. (1997). Attenuating verbal overshadowing through color retrieval cues. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 23*(4), 915–931. https://doi.org/10.1037/0278-7393.23.4.915.

Brown, C., Brandimonte, M. A., Wickham, L. H. V., Bosco, A., & Schooler, J. W. (2014). When do words hurt? A multiprocess view of the effects of verbalization on visual memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 40*(5), 1244–1256. https://doi.org/10.1037/a0037222.

Carmichael, L., Hogan, H. P., & Walter, A. A. (1932). An experimental study of the effect of language on the reproduction of visually perceived form. *Journal of Experimental Psychology, 15*(1), 73–86. https://doi.org/10.1037/h0072671.

Cibelli, E., Xu, Y., Austerweil, J. L., Griffiths, T. L., & Regier, T. (2016). The Sapir-Whorf Hypothesis and Probabilistic Inference: Evidence from the Domain of Color. *PLOS ONE, 11*(7), Article e0158725. https://doi.org/10.1371/journal.pone.0158725.

Crawford, L. E., Huttenlocher, J., & Engebretson, P. H. (2000). Category effects on estimates of stimuli: Perception or reconstruction? *Psychological Science, 11*(4), 280–284. https://doi.org/10.1111/1467-9280.00256.

Donkin, C., Nosofsky, R., Gold, J., & Shiffrin, R. (2015). Verbal labeling, gradual decay, and sudden death in visual short-term memory. *Psychonomic Bulletin & Review, 22*(1), 170–178. https://doi.org/10.3758/s13423-014-0675-5.

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.

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), 1105–1123. https://doi.org/10.1037/xge0000560.

Forsberg, A., Johnson, W., & Logie, R. H. (2020). Cognitive aging and verbal labeling in continuous visual memory. *Memory & Cognition, 48*(7), 1196–1213. https://doi.org/10.3758/s13421-020-01043-3.

Franklin, A., Drivonikou, G. V., Bevis, L., Davies, I. R. L., Kay, P., & Regier, T. (2008). Categorical perception of color is lateralized to the right hemisphere in infants, but to the left hemisphere in adults. *Proceedings of the National Academy of Sciences, 105*(9), 3221–3225. https://doi.org/10.1073/pnas.0712286105.

Gilbert, A. L., Regier, T., Kay, P., & Ivry, R. B. (2006). Whorf hypothesis is supported in the right visual field but not the left. *Proceedings of the National Academy of Sciences of the United States of America, 103*(2), 489–494. https://doi.org/10.1073/pnas.0509868103.

Gobet, F., Lane, P. C. R., Croker, S., Cheng, P. C.-H., Jones, G., Oliver, I., & Pine, J. M. (2001). Chunking mechanisms in human learning. *Trends in Cognitive Sciences, 5*(6), 236–243. https://doi.org/10.1016/S1364-6613(00)01662-4.

Hanley, J. R., & Roberson, D. (2011). Categorical perception effects reflect differences in typicality on within-category trials. *Psychonomic Bulletin & Review, 18*(2), 355–363. https://doi.org/10.3758/s13423-010-0043-z.

Hardman, K. O., Vergauwe, E., & Ricker, T. J. (2017). Categorical working memory representations are used in delayed estimation of continuous colors. *Journal of Experimental Psychology: Human Perception and Performance, 43*(1), 30–54. https://doi.org/10.1037/xhp0000290.

Hardman, K. O. (2016). CatContModel: Categorical and Continuous working memory models for delayed estimation tasks (Version 0.7.1) [Computer software]. https://github.com/hardmanko/ CatContModel/releases/tag/v0.6.1.

Huang, L., & Awh, E. (2018). Chunking in working memory via content-free labels. *Scientific Reports, 8*(1), 23. https://doi.org/10.1038/s41598-017-18157-5.

Huttenlocher, J., Hedges, L. V., & Duncan, S. (1991). Categories and particulars: Prototype effects in estimating spatial location. *Psychological Review, 98*(3), 352–376. https://doi.org/10.1037/0033-295X.98.3.352.

Kelly, L. J., & Heit, E. (2017). Recognition memory for hue: Prototypical bias and the role of labeling. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 43*(6), 955–971. https://doi.org/10.1037/xlm0000357.

Kruschke, J. K. (2013). Bayesian estimation supersedes the t test. *Journal of Experimental Psychology: General, 142*(2), 573–603. https://doi.org/10.1037/a0029146.

Lewandowsky, S., & Oberauer, K. (2015). Rehearsal in serial recall: An unworkable solution to the nonexistent problem of decay. *Psychological Review, 122*(4), 674–699. https://doi.org/10.1037/a0039684.

Li, A. Y., Liang, J. C., Lee, A. C. H., & Barense, M. D. (2020). The validated circular shape space: Quantifying the visual similarity of shape. *Journal of Experimental Psychology: General, 149*(5), 949–966. https://doi.org/10.1037/xge0000693.

Luck, S. J., & Vogel, E. K. (2013). Visual working memory capacity: From psychophysics and neurobiology to individual differences. *Trends in Cognitive Sciences, 17*(8), 391–400. https://doi.org/10.1016/j.tics.2013.06.006.

Lupyan, G. (2008). 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. (2012a). Linguistically modulated perception and cognition: The label-feedback hypothesis. *Frontiers in Psychology, 3*, 54.

Lupyan, G. (2012b). What do words do? Toward a theory of language-augumented thought. *Psychology of Learning and Motivation, 57*, 255–297.

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., & 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., & 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.

Meissner, C. A., Sporer, S. L., & Susa, K. J. (2008). A theoretical and meta-analytic review of the relationship between verbal descriptions and identification accuracy in memory for faces. *European Journal of Cognitive Psychology, 20*(3). http://works.bepress.com/christian_meissner/32/.

Morey, R. D., & Rouder, J. N. (2015). BayesFactor: Computation of Bayes factors for common designs (0.9.12-2) [Computer software]. http://CRAN.R-project.org/package=BayesFactor.

Morey, C. C., & Cowan, N. (2004). When visual and verbal memories compete: Evidence of cross-domain limits in working memory. *Psychonomic Bulletin & Review, 11*(2), 296–301. https://doi.org/10.3758/BF03196573.

Morey, C. C., & Cowan, N. (2005). When do visual and verbal memories conflict? The importance of working-memory load and retrieval. *Journal of Experimental Psychology. Learning, Memory, and Cognition, 31*(4), 703–713. https://doi.org/10.1037/0278-7393.31.4.703.

Murdock, B. (1960). The distinctiveness of stimuli. *Psychological Review, 67*(1), 16–31. https://doi.org/10.1037/h0042382.

Nassar, M. R., Helmers, J. C., & Frank, M. J. (2018). Chunking as a rational strategy for lossy data compression in visual working memory. *Psychological Review, 125*(4), 486–511. https://doi.org/10.1037/rev0000101.

Oberauer, K. (2019). Is rehearsal an effective maintenance strategy for working memory? *Trends in Cognitive Sciences*. https://doi.org/10.1016/j.tics.2019.06.002.

Oberauer, K., Stoneking, C., Wabersich, D., & Lin, H.-Y. (2017). Hierarchical Bayesian measurement models for continuous reproduction of visual features from working memory. *Journal of Vision, 17*(5), 11. https://doi.org/10.1167/17.5.11.

Pansky, A., & Koriat, A. (2004). The Basic-Level Convergence Effect in Memory Distortions. *Psychological Science, 15*(1), 52–59. https://doi.org/10.1111/j.0963-7214.2004.01501009.x.

Pavio, A. (1991). Dual coding theory: Retrospect and current status. *Canadian Journal of Psychology/Revue Canadienne de Psychologie, 45*(3), 255–287.

Pelli, D. G. (1997). The VideoToolbox software for visual psychophysics: Transforming numbers into movies. *Spatial Vision, 10*(4), 437–442. https://doi.org/10.1163/156856897X00366.

Persaud, K., & Hemmer, P. (2016). The dynamics of fidelity over the time course of long-term memory. *Cognitive Psychology, 88*, 1–21. https://doi.org/10.1016/j.cogpsych.2016.05.003.

Pratte, M. S., Park, Y. E., Rademaker, R. L., & Tong, F. (2017). Accounting for stimulus-specific variation in precision reveals a discrete capacity limit in visual working memory. *Journal of Experimental Psychology: Human Perception and Performance, 43*(1), 6–17. https://doi.org/10.1037/xhp0000302.

Prinzmetal, W., Amiri, H., Allen, K., & Edwards, T. (1998). Phenomenology of attention: I. Color, location, orientation, and spatial frequency. *Journal of Experimental Psychology: Human Perception and Performance, 24*(1), 261–282. https://doi.org/10.1037/0096-1523.24.1.261.

R core team. (2017). R: A language and environment for statistical computing. R Foundation for Statistical Computing. http://www.R-project.org/.

Regier, T., Kay, P., & Khetarpal, N. (2007). Color naming reflects optimal partitions of color space. *Proceedings of the National Academy of Sciences, 104*(4), 1436–1441. https://doi.org/10.1073/pnas.0610341104.

Richler, J. J., Palmeri, T. J., & Gauthier, I. (2013). How does using object names influence visual recognition memory? *Journal of Memory and Language, 68*(1), 10–25. https://doi.org/10.1016/j.jml.2012.09.001.

Ricker, T. J. (2015). The role of short-term consolidation in memory persistence. *AIMS Neuroscience, 2*(4), 259–279. https://doi.org/10.3934/Neuroscience.2015.4.259.

Roberson, D., & Davidoff, J. (2000). The categorical perception of colors and facial expressions: The effect of verbal interference. *Memory & Cognition, 28*(6), 977–986. https://doi.org/10.3758/BF03209345.

Rouder, J. N. (2014). Optional stopping: No problem for Bayesians. *Psychonomic Bulletin & Review, 21*(2), 301–308. https://doi.org/10.3758/s13423-014-0595-4.

Schönbrodt, F. D., Wagenmakers, E.-J., Zehetleitner, M., & Perugini, M. (2017). Sequential hypothesis testing with Bayes factors: Efficiently testing mean differences. *Psychological Methods, 22*(2), 322–339. https://doi.org/10.1037/met0000061.

Schooler, J. W., & Engstler-Schooler, T. Y. (1990). Verbal overshadowing of visual memories: Some things are better left unsaid. *Cognitive Psychology, 22*(1), 36–71.

Sense, F., Morey, C. C., Prince, M., Heathcote, A., & Morey, R. D. (2016). Opportunity for verbalization does not improve visual change detection performance: A state-trace analysis. *Behavior Research Methods, 1–10*. https://doi.org/10.3758/s13428-016-0741-1.

Skelton, A. E., Catchpole, G., Abbott, J. T., Bosten, J. M., & Franklin, A. (2017). Biological origins of color categorization. *Proceedings of the National Academy of Sciences, 114*(21), 5545–5550. https://doi.org/10.1073/pnas.1612881114.

Sloutsky, V. M., & Fisher, A. V. (2012). Linguistic labels: Conceptual markers or object features? *Journal of Experimental Child Psychology, 111*(1), 65–86. https://doi.org/10.1016/j.jecp.2011.07.007.

Souza, A. S., & Oberauer, K. (2018). Does articulatory rehearsal help immediate serial recall? *Cognitive Psychology, 107*, 1–21. https://doi.org/10.1016/j.cogpsych.2018.09.002.

Souza, A. S., & Oberauer, K. (2020). No evidence that articulatory rehearsal improves complex span performance. *Journal of Cognition, 3*(1). https://doi.org/10.5334/joc.103.

Souza, A. S., Rerko, L., Lin, H.-Y., & Oberauer, K. (2014). Focused attention improves working memory: Implications for flexible-resource and discrete-capacity models. *Attention, Perception, & Psychophysics, 76*(7), 2080–2102. https://doi.org/10.3758/s13414-014-0687-2.

Souza, A. S., & Skóra, Z. (2017). The interplay of language and visual perception in working memory. *Cognition, 166*, 277–297. https://doi.org/10.1016/j.cognition.2017.05.038.

Thierry, G., Athanasopoulos, P., Wiggett, A., Dering, B., & Kuipers, J.-R. (2009). Unconscious effects of language-specific terminology on preattentive color perception. *Proceedings of the National Academy of Sciences, 106*(11), 4567–4570. https://doi.org/10.1073/pnas.0811155106.

Uchikawa, K., & Shinoda, H. (1996). Influence of basic color categories on color memory discrimination. *Color Research & Application, 21*(6), 430–439. https://doi.org/10.1002/(SICI)1520-6378(199612)21:6<430::AID-COL5>3.0.CO;2-X.

van den Berg, R., Awh, E., & Ma, W. J. (2014). Factorial comparison of working memory models. *Psychological Review, 121*(1), 124–149. https://doi.org/10.1037/a0035234.

Vogel, E. K., Woodman, G. F., & Luck, S. J. (2001). Storage of features, conjunctions, and objects in visual working memory. *Journal of Experimental Psychology: Human Perception and Performance, 27*(1), 92–114. https://doi.org/10.1037/0096-1523.27.1.92.

Wilken, P., & Ma, W. J. (2004). A detection theory account of change detection. *Journal of Vision, 4*(12), 11. https://doi.org/10.1167/4.12.11.

Winawer, J., Witthoft, N., Frank, M. C., Wu, L., Wade, A. R., & Boroditsky, L. (2007). Russian blues reveal effects of language on color discrimination. *Proceedings of the National Academy of Sciences, 104*(19), 7780–7785. https://doi.org/10.1073/pnas.0701644104.

Zentall, T. R., Wasserman, E. A., Lazareva, O. F., Thompson, R. K. R., & Rattermann, M. J. (2008). Concept learning in animals. *Comparative Cognition & Behavior Reviews, 3*, 13–45. https://doi.org/10.3819/ccbr.2008.30002.

Zhang, W., & Luck, S. J. (2008). Discrete fixed-resolution representations in visual working memory. *Nature, 453*(7192), 233–235. https://doi.org/10.1038/nature06860.