

Recognition Memory for Hue: Prototypical Bias and the Role of Labeling

Laura Jane Kelly and Evan Heit
University of California, Merced

How does the concurrent use of language affect perception and memory for exemplars? Labels cue more general category information than a specific exemplar. Applying labels can affect the resulting memory for an exemplar. Here 3 alternative hypotheses are proposed for the role of labeling an exemplar at encoding: (a) labels distort memory toward the label prototype, (b) labels guide the level of specificity needed in the current context, and (c) labels direct attention to the label's referent among all possible features within a visual scene. University students were shown hues on object silhouettes that they either labeled with basic color categories, made preference judgments about, or indicated the animacy of its category. Experiments 1 and 2 established that there are response shifts toward the category prototype regardless of labeling, showing a pervasive influence of category knowledge on response bias. They also established an effect of labeling whereby labeling decreases the magnitude of shifts. Experiments 3 and 4 investigated the uniqueness and necessity of language in causing the decreased shift—neither of which proved to be the case. Overall, category-relative bias was pervasive and labeling appears to direct attention to the feature resulting in less biased memory. The results highlight that the context at encoding affects how memory is formed.

Keywords: labeling, categorization, recognition memory, memory bias

What are the consequences of labeling for perception and memory? Traditionally, language was treated as a separate system within human cognition as it is uniquely human, whereas memory, attention, and perception are held in common with evolutionary ancestors (Chomsky, 1975; Fodor, 1983; Pinker, 1994). However, there is increasing evidence that language uses the same underlying subsystems as other cognitive processes (Perfors, Tenenbaum, & Regier, 2011; Tomasello, 2009). Within such accounts of cognitive processing, interaction between language and memory no longer needs to be the product of separate processes taking place concurrently. Rather, the processes themselves may overlap in their implementation with a high degree of mutual influence.

Linguistic knowledge and nonlinguistic knowledge are intimately tied together. To communicate about knowledge, the most pervasive and efficient form of information transfer is through language. Categories are structured knowledge about groups of objects, actions, or whatever phenomena are able to meaningfully hang together. In everyday behavior, category labels are used to refer to phenomena in the world as people communicate with one another and themselves. Many successful models of categorization do not assign any unique significance to the act of applying

category labels (Heit, 1992; Nosofsky, 2015). The topic has started to garner considerable interest (Blanco & Gureckis, 2013; Hanley & Roberson, 2011; Lupyan, 2008; Lupyan, Rakison, & McClelland, 2007; Richler, Gauthier, & Palmeri, 2011; Richler, Palmeri, & Gauthier, 2013; Roberson & Hanley, 2010). If the act of labeling systematically affects concurrent perception and subsequent memory, it could have pervasive influence on knowledge built out of our experience with the world.

In the present paper we address the issue of how labeling a perceived hue, for example, calling an orange-red hue “red,” affects subsequent memory for that hue, particularly whether there are systematic shifts in memory responses relative to the category typicality structure. We compare labeling to other judgments made at encoding in order to examine the specific effects of labeling on memory for percepts and to examine whether these effects are unique to labeling.

Memory Distortion

Memory is remarkable in its capacity. People are capable of distinctly remembering hundreds of objects (Brady, Konkle, Alvarez, & Oliva, 2008) and scenes (Konkle, Brady, Alvarez, & Oliva, 2010; Shepard, 1967) after relatively brief exposure. At the same time, memory is not necessarily accurate to experience even when a person is very confident of that memory. After hearing an intense verb to describe a car crash—*smashed*—participants reported remembering broken glass—a marker of a more intense crash—even though there was no broken glass in the video of the crash they had watched (Loftus & Palmer, 1974). Whereas memory is clearly not analogous to video recordings capturing unaltered snippets of audiovisual reality, documented distortions like the eyewitness testimony example have systematic structure. A false alarm—remembering something as old when it is new—in memory for a list of words is more likely if the word is semanti-

This article was published Online First January 12, 2017.

Laura Jane Kelly and Evan Heit, Cognitive and Information Sciences, University of California, Merced.

This material is based upon work while Evan Heit was serving at the National Science Foundation (US). Any opinion, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

Correspondence concerning this article should be addressed to Laura Jane Kelly, School of Social Sciences, Humanities, and Arts, University of California, Merced, 5200 North Lake Road, Merced, CA 95343. E-mail: lkelly@ucmerced.edu

cally related to the rest of the words in the list (Roediger & McDermott, 1995). The word habitually used to describe the middle color of a traffic light (yellow—*gelb*—in German and orange—*oranje*—in Dutch) influences the memory of the light's hue (Mitterer, Horschig, Musseler, & Majid, 2009). In situations that are selected for their contrast between reality and description, language is able to influence memory.

How is memory affected by everyday application of language during perception? In situations where language and perception match, does language affect memory for that perception? For instance, as we examine in the present experiments, does labeling a specific hue as “red” change the memory for that hue? A specific hue and the general category of red do not contain the same information. The label red applies to the hue but it also applies to a range of other similar hues. The mismatch between information the label can apply to and the specific item the label is being applied to at a particular point in time could potentially lead to systematic memory distortions. However, that a label can apply to a broad range of category members does not necessarily mean the label in context activates category information with enough strength to alter processing of perceptual input.

How Could Labeling Affect Perception and Memory?

If there were a pervasive influence of labeling on perception and memory, even when the labels agree with the perceptual input, what would this influence be? Here, we explore three alternative hypotheses: (a) labels distort memory toward the label prototype, (b) labels guide the level of specificity needed in the current context, and (c) labels direct attention to the label's referent among all possible features within a visual scene.

Labels Distorting Memory

The representational shift hypothesis (Lupyan, 2008) makes the conjecture that memory is systematically biased toward the category prototype when category labels are active at the same time as when relevant perception is taking place. Specifically, the representational shift hypothesis suggests that perceptual information is perceived equivalently regardless of labeling; then, while the information is processed, there is memory distortion from top-down category information being coactivated with the bottom-up perceptual information. The category information is broader than the specific exemplar being perceived. Labeling causes the category information to be more strongly activated than if the category was activated through the perceptual information alone. The general category information and the specific exemplar interact in online processing, shifting the memory of the original percept toward the prototype of the category—the overlapping information between percept and category is more active than idiosyncratic aspects of the current exemplar. It is that shifted representation that is stored in long-term memory. In other words, the orange-red hue labeled as red would be remembered as more red. However, this pervasive prototypical distortion of memory based on labeling has been questioned (Blanco & Gureckis, 2013; Richler et al., 2011, 2013). The representational shift hypothesis, if correct, would represent an accumulating influence of existing knowledge biasing new knowledge toward existing ideas about how the world is structured.

Labels Guiding Specificity

Another possible effect is that the label serves as a cue as to what level of encoding specificity is needed. For example, the very question of whether a hue is red or green may suggest that the diagnostic level of detail needed to distinguish the hues is all that is needed in the current environment, rather than more detailed information such as whether it is an orange-tinted red or more of a purple-tinted red. In this view, an item only needs to be processed sufficiently to be confident the item is in one category and not the other when choosing between two labels, as was the case in related work (Richler et al., 2011). A superficial orienting task, such as indicating the direction a chair is facing (Blanco & Gureckis, 2013) or the location of the item on the display screen (Richler et al., 2011), could also lead to poor memory for the exemplar. These tasks were designed to influence the strength of a memory through tasks requiring low depth of processing. Under this account, labeling does not have a systematic influence on memory traces so much as simply being a task that may not require specificity.

Labels as an Attentional Focus

More broadly, labels may serve as a general cue for attention. When a stimulus or a scene is visually perceived, there are many features initially processed into iconic memory. Guiding the focus of attention to regions within a scene guides the content of accessible working memory for that scene. This was demonstrated in a classic study (Sperling, 1960) by briefly showing participants a display of many rows and columns then immediately cueing a specific row. Participants were able to remember any single row that was cued, but they could not remember the whole array. Focusing attention to a particular row allowed that information to be secured into memory while the perception of the rest of the display was lost. Likewise, using a label concurrently with visual perception would focus the strength of memory encoding toward the features relevant to the label. Under this account, labeling a percept in terms of its color would encourage attention to featural information related to color, but labeling would not be unique in this role, and it should be possible to cue color in other ways as well.

Existing Evidence for Systematic Bias

Labeling bias as discussed here has investigated in several recent studies using an old/new recognition paradigm (Blanco & Gureckis, 2013; Lupyan, 2008; Richler et al., 2011). A set of items is presented during a study block, during which time participants are (or are not) asked to label each item. Next, these items and matched lures are presented during a test block. There are challenges associated with using this paradigm to test the representational shift hypothesis, which predicts that memory is shifted systematically toward the prototype (Lupyan, 2008). Poor recognition memory, in the form of misses on an old/new recognition test, need not mean that memory for that item is more prototypical. The memory could be more prototypical, less prototypical, or simply not strong enough to produce recognition regardless of movement in representational space. Returning to our earlier example, if a participant studies an orange-red hue, a failure to later

recognize that hue may or may not mean that the representation had shifted toward the red prototype.

In our own prior work (Kelly & Heit, 2014), we had focused on the claim (Lupyan, 2008) that memory shifts during the encoding process. Therefore, we tested immediate memory with narrow study-to-test timescales, 500 ms and 5,000 ms. Memory for hue was tested using the old study items as well as four matched lures that systematically varied relative to the color prototype. It was the responses on the lure items, either shifted toward or away from the prototype, that were crucial for examining whether representation had shifted and how. We found that at these timescales, memory test responses were systematically biased relative to the category structure and, surprisingly, are biased away from the prototype.

These shift patterns may not hold at other timescales. It seems possible that, at a timescale of minutes, memory response distortion during recognition would be prototypical rather than atypical. In fact, much of what is known of memory including gist memory (Reyna & Brainerd, 1995) and false recognition of close associates in word lists (Roediger & McDermott, 1995) suggests that memory is broadly influenced by category knowledge and that influence is generally in a prototypical direction.

Additionally, at the brief timescales investigated in Kelly and Heit (2014), there was not a distinctive effect of language. There was a small effect of category labeling that emerged by 5,000 ms but was not present at 500 ms. Encoding hue into immediate memory was not apparently affected by online labeling. To determine if online labeling influences processing of hue over time, particularly as memory traces get encoded into long-term memory, here we used a hybrid of our earlier method with a longer study-to-test delay based on the timing used by other researchers (Blanco & Gureckis, 2013; Lupyan, 2008; Richler et al., 2011).

In related work, Bae, Olkkonen, Allred, and Flombaum (2015) investigated hue estimation (rather than recognition) at 0 ms and 900 ms, finding varying amounts and direction of bias depending on target hue. However, the Bae et al. experiments did not have a labeling manipulation. Also, Persaud and Hemmer (2014) looked at bias in color recall memory using active labeling immediately before recall, estimating the hue on a color wheel. They found memory to be biased toward the prototype or mean hue value. However, there was no comparison condition without labeling, and labeling itself occurred just before the response. Hence neither of these studies allows us to address questions about the influence of labeling at encoding and whether an effect found is unique to labeling.

Shifts of What Exactly?

The key issue of whether memory truly shifts, and, if it does shift, how it shifts relative to the prototype is not discernible from an old/new recognition paradigm with a single matched lure as described above. A different paradigm, one that probes memory at multiple locations, can shed some light on whether the hypothesized shifts are a possibility. If memory test responses are systematically shifted, the role of language in causing those shifts then becomes more interesting. Even showing shifts in memory test responses does not demonstrate that memory itself is shifted—it could be that shifts are introduced at retrieval rather than encoding. However, the mechanism debate is premature without first demonstrating shifts in memory responses.

Yet, mechanisms of shifts in addition to whether shifts exist have already been debated. The representational shift hypothesis (Lupyan, 2008) describes a phenomenon that is specific to overt labeling of percepts. It does not predict that merely knowing a category will cause prototypical shifts of exemplar memory, but that using the label of the category actively changes the processing of perceptual input. This leads to the following prediction: Labeled images will be stored more prototypically and the original image will therefore be less likely to be recognized during a memory test. This would be due to a change in the encoded percept. The specificity of encoding view predicts a different mechanism: less specificity in the memory trace and therefore less recognition of the original labeled images (Blanco & Gureckis, 2013; Richler et al., 2011). The best guess would be a more prototypical guess. Memory probed at multiple locations on a typicality gradient would be predicted to have the same result by both these accounts: labeled items are expected to be more prototypical than unlabeled items. The directed attention account would, in opposition to the other accounts, predict less shifting for labeled items than unlabeled items. This decrease in shifts would be due to more specificity in encoding due to the key aspect of the item being highlighted when perceived. Shifts in the response for the second and third hypotheses would not necessarily be shifts of the actual percept, but could reflect relative susceptibility of a memory trace to the influence of a category prototype at response. For example, a potential mechanism could be that described by a prominent model of spatial memory, the category adjustment model (CAM; Huttenlocher, Hedges, & Duncan, 1991; Huttenlocher, Hedges, Lourenco, Crawford, & Corrigan, 2007; Huttenlocher, Hedges, & Vevea, 2000). CAM predicts shifts in spatial memory based on the strength of a memory trace with less specific exemplar memory filled in with category information at retrieval. Applying the principles of CAM to hue memory, memory traces would be encoded as originally perceived but less specific memories should result in more shifted responses while more specific memories would result in less shifted responses.

Experiment Rationale

We conducted a series of four experiments using a hybrid of paradigms from Lupyan (2008) and Kelly and Heit (2014). The aim of the work was to investigate whether there is a systematic bias predicated on labeling categories when encountering a percept. To give a preview, participants were presented with colored silhouettes and made two types of judgments about them (with different pairs of judgments in each experiment; Table 1). Afterward they were given a five-alternative forced-choice memory test for the hue of the silhouette among slightly different hues: an example display is given in Figure 1. Experiments 1 and 2 established that shifts were taking place. Additionally, the experiments investigated whether overt labeling of categories leads to a different pattern of shifts than a comparison condition: in Experiment 1 with the preference judgment comparison condition as in Lupyan (2008), then in Experiment 2 with another categorization task, an animacy comparison condition. Experiments 3 and 4 explored the role of language in shifts, particularly in Experiment 3 whether the effect of language is unique and in Experiment 4 whether overt color labeling is required to produce the effect. This sequence of experiments allowed us to test for systematic shifts relative to

Table 1
Proportions of Responses by Typicality Level at Test

Experiment	Condition	Least typical	Less typical	Original	More typical	Most typical
Exp. 1	Color	.155	.155	.191	.236	.263
	Preference	.127	.159	.173	.234	.307
Exp. 2	Color	.170	.163	.171	.223	.273
	Animacy	.125	.143	.177	.202	.354
Exp. 3	Color	.182	.132	.177	.208	.302
	Animacy	.192	.143	.150	.240	.275
Exp. 4	Preference	.170	.127	.181	.192	.330
	Animacy	.165	.146	.180	.208	.301

category structure, to determine the direction of those shifts, and to examine the role of online labeling in producing shifts at the timescale of about five minutes between study and test.

Experiment 1

To look for shifts in memory responses after a few minutes of processing, we employed the basic design of the original representational shift paradigm (Lupyan, 2008): a study phase during which participants viewed target items and made categorization or preference judgments about them, immediately followed by a surprise test phase where participants identified which items they had seen previously. Rather than display targets and a single matched foil in isolation at test, we tested participants on an array of five hues: the original hue along with two hues more typical and two hues less typical of the color category. The richer false alarm possibilities make the direction and magnitude of a response shift observable.

In line with Kelly and Heit (2014), we used color categories rather than object categories such as chairs and lamps used in the studies by other researchers because color has a better defined representational space. We focused on the single dimension of hue, leaving two directions of possible shift: away from or toward the prototype of a color category. Color is a domain where memory has, in some cases, proven accurate to experience (Pérez-Carpinell, Baldoví, de Fez, & Castro, 1998), and is a domain where categories affect perception (Kay & Kempton, 1984; Winawer et al., 2007). The possible veracity of perception coupled with the demonstrated distortion of perception in different settings make color an ideal testing ground for whether overt language can cause systematic memory distortion.

Experiment 1 began to address both of our aims: establishing whether there are directional shifts in memory test responses and, should shifts be found, investigating the role of language in those shifts.

Method

Participants. Thirty-eight undergraduate students were recruited from University of California, Merced (UC Merced)'s participant pool.¹ All participants were monolingual or early (by age 10) bilingual English speakers. They had not otherwise participated in color-based experiments in our lab. Participants had normal or corrected to normal vision. They were also tested to

have normal color vision, using the CITY test (City University, 2002) at the conclusion of the experimental session.

Stimuli. Forty silhouettes were created in Adobe Photoshop for the study phase. Twenty silhouettes were living things such as a giraffe and a butterfly, and 20 silhouettes were nonliving things such as a pan and an airplane. Eight main target hues (4 reds and 4 greens) were selected. The selected colors had the lightness and saturation values of their category's focal color (Sturges & Whitfield, 1995). The hues were evenly distanced from each other within hue space centered on the category focal color hue. The semirandom (see Appendix A for details) creation of colored silhouettes preserved color and animacy balances while counterbalancing the color/shape pairings across participants. The animacy features of the silhouettes were not required in the current experiment but were necessary for an animacy judgment condition used in Experiments 2–4. Six additional colored silhouettes were developed for practice trials, using the colors yellow and purple as well as different shapes than those used in the main task.

For the test phase, four variations of each of the eight target hues were created at intervals of 2° along the hue dimension in CIE L*a*b color space (see Appendix A for details). This resulted in an evenly spaced test scale for each silhouette of two hues closer to the prototype: the original hue and two hues further from the prototype. The typicality of the hues was empirically tested in a norming experiment using an independent participant sample, which is reported in Appendix B.

Procedure. The experiment consisted of two main parts: a study phase and a test phase. Each participant encountered six practice study trials, then 80 study trials followed by 40 test trials.

For the study phase, participants were instructed in the two types of judgments: color categorization, "What color is this?" ("1" and "2" for red and green, respectively, with color/key assignments counterbalanced across participants) and preference, "Do you like this?" ("3" for like or "4" for dislike), the conditions used in Lupyan (2008). The trials were presented in eight alternating judgment blocks (e.g., a block of red/green, followed by a block of like/dislike), 10 trials per block, with the starting condition counterbalanced. Each silhouette was judged twice by each participant for the same judgment type (e.g., if the giraffe was judged for color the first time it appeared, the giraffe would again be judged for color the second time it appeared). Each silhouette was judged in each condition, between participants. The participants had been instructed to remember the silhouettes as they would show up more than once but were not explicitly told of a memory test. Each trial consisted of a fixation cross (1,500 ms), the silhouette to be judged (300 ms), a question mark eliciting the judgment for that block (up to 700 ms), and a blank screen (up to 1,000 ms). The trial ended after a response was given. If the trial timed out a "Too Slow" message briefly appeared on the screen before a new trial started. The timing and block design of the study phase closely followed the procedure of Lupyan's (2008) Experiment 1.

After the study phase, participants were then tested for hue memory. The memory trials consisted of a circular array of the five hue variants of a particular silhouette (see Figure 1). The array

¹ A preliminary version of this experiment was presented in a conference proceedings paper (Kelly & Heit, 2013). We have omitted one participant due to missing information on native-English speaking status.

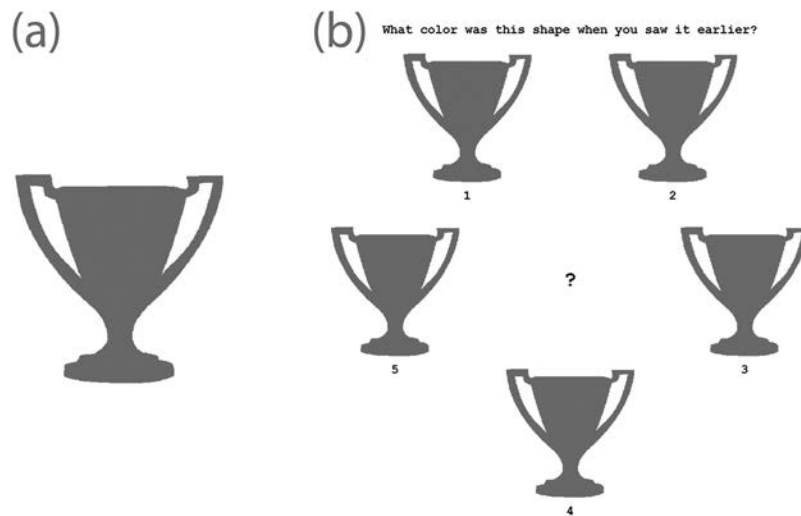


Figure 1. Stimulus presentation. (a) During the study phase, participants saw a single silhouette as they made a judgment. (b) During the test phase, participants selected from an array of five hues for a particular silhouette shape. The question reads, “What color was this shape when you saw it earlier?” See the online article for the color version of this figure.

consisted of the hues in graded clockwise order with the most typical hue rotated to a random position on each trial, resulting in a consistent appearance of selecting from a gradient of hues but avoiding position effects that would be present in a line. Each of the five positions had a location label 1 through 5 that participants entered on the keyboard to make a selection. There was no time limit imposed on the memory test responses, with an intertrial interval of 1,500 ms.

Results and Discussion

In spite of screening for color vision deficiencies, four participants were excluded for low (<80%) color categorization accuracy. Six additional participants were excluded for response patterns suggesting a lack of engagement such as always pressing the same response key. As a result, 28 participants were included in the analyses. The same pattern of results was maintained when all 38 participants were analyzed. The included participants had a color categorization accuracy of 90.56%.²

The first consideration was whether there were the hypothesized systematic shifts. Responses on the memory test were transformed to a consistent -2 to $+2$ scale, with -2 corresponding to choosing the item 4° of hue space more distant from the prototype than the original, 0 choosing the original item, and $+2$ choosing the item 4° closer to the prototype than the original. The proportion of responses for each of these hue types by condition are reported in Table 1.

Participants showed a systematically biased memory for hue toward the prototype. We conducted two-tailed, one-sample t tests comparing the mean typicality of responses for each judgment type against the value 0 representing the original hues. The color categorization mean was 0.29 (standard deviation 0.42) which is significantly more prototypical than the original hues, $t(27) = 3.75$, $p < .001$, Cohen's $d = 0.71$, $BF_{10} = 38.97$. Here and in subsequent reports, the Bayes factor (BF), calculated using JASP

(Love et al., 2015), represents the evidence of the alternative hypothesis (H_1) against the null hypothesis (H_0), with values above 1 being evidence for H_1 and values between 1 and 0 being evidence for H_0 . To interpret the current value, a BF_{10} of 30–100 is considered very strong support for the alternative hypothesis (Jeffreys, 1961). The preference judgment condition mean was 0.44 (standard deviation of 0.42) which is also significantly more prototypical than the original hues, $t(27) = 5.57$, $p < .001$, $d = 1.05$, $BF_{10} > 1,000$. A Bayes factor above 100 is considered decisive support for the alternative hypothesis. Here, we provide direct evidence for shifts toward the prototype at a study-to-test delay of approximately 5 min. In an effort to verify the assumptions of these analyses, we examined the response distributions for both color judgments and preference judgments. Each distribution failed a test for nonnormality, the Shapiro-Wilk test, p values $> .05$. The distributions were each not statistically different from a normal distribution.

This design did not have a measure of hue/shape pair memory. The memory test does not distinguish random guessing due to no memory for the hue when a shape is displayed at test and true biased memory. We conducted a control experiment reported in Appendix B. Participants in that experiment had a 67% accuracy rate of categorizing a gray silhouette cue as having been red or green at study—effectively a cued recall task. Participants clearly have some memory of the pairings but it is not at ceiling. To determine whether there were differences in accurate hue memory across the encoding tasks, we analyzed the proportion of objectively correct responses (the Original column in Table 1) by condition. There was no difference in these proportions (paired, two-tailed $t(27) = 0.92$, $p > .05$, $d = 0.17$, $BF_{10} = 0.29$), indicating that the conditions contained equal original hue re-

² No individual hues were more accurately categorized than the others (see Appendix A for details).

sponses with any differences between conditions being within the pattern of the lure responses. Across all four experiments, the findings illustrate that simply looking at these accurate responses to the original hue is not as informative as looking at the whole pattern of false alarms.

In order to investigate potential condition differences, we tested the mean response typicality of the labeling and preference conditions against each other using a two-tailed, paired-sample *t* test. The difference between conditions (label condition mean: 0.29 and preference condition mean: 0.44) was not significant, $t(27) = 1.97$, $p = .059$, $d = 0.34$, $BF_{10} = 1.08$. With a *BF* just above 1, a difference between the conditions is only weakly evident. The null hypothesis of no difference is not supported, leaving the condition difference inconclusive. A chi-square analysis comparing the distribution of responses at each typicality level between the two conditions was also nonsignificant, $\chi^2(4, N = 1,160) = 3.51$, $p > .05$. Since the *t* test yielded a *p* value close to the significance level of .05 and an inconclusive *BF*, it is natural to speculate as to whether there is a difference that we do not have enough power to detect. However, note that the shift was actually weaker for the labeling condition, the opposite pattern of results from those proposed by the representational shift hypothesis (Lupyan, 2008). If this pattern holds, color labeling would actually lead to weaker rather than stronger shifts in comparison to another kind of judgment. Moreover, labeling does not uniquely cause shifted response patterns. This cannot support the specific formulation of memory distortion hypothesized by the representational shift hypothesis. The other possibilities of labels guiding specificity or focusing attention could be compatible with the demonstrated systematic bias.

Finally, we collected reaction times (RTs) for both judgment responses and memory test responses with measurement beginning at stimulus presentation—either the single item in the study phase or the array of items in the test phase. The color judgment RTs (median 425.62 ms) were significantly faster than the preference judgment RTs (median 520.10 ms) according to the Wilcoxon signed-ranks test ($V = 71$, $p < .002$) which is appropriate for comparing paired sample nonnormal distributions. This replicates an RT difference reported in Lupyan (2008). The additional processing time needed for preference judgments has been taken as indicative of more depth of processing in the preference condition (Blanco & Gureckis, 2013; Lupyan, 2008; Richler et al., 2011, 2013). There was no significant difference for color memory RTs (median 3,444.20 ms) and preference memory RTs (median 3,982.33 ms), $V = 185$, $p = .695$.

Overall, the main finding of Experiment 1 was that responses shift relative to category structure at the retention interval of a few minutes. This happened regardless of the orienting task being color labeling or preference judgment. Additionally, there was not a significant difference between the two orienting conditions. The pattern of results, though not conclusive, lean toward *less* bias for labeled hues, the opposite pattern from what the representational shift hypothesis would predict. The possible roles of labels setting a minimum level of specificity and of labels focusing attention to particular features are not distinguished by Experiment 1. Preference judgments are not categorical and specific in the same way as making category label judgments are. The level of specificity needed to make a preference judgment is less clear than discriminating between two well-defined categories in the color label

condition. Which features preference judgments cause a participant to focus on is unclear and likely subject to individual differences with some participants focusing on liking or disliking the color, the shape, the animal or object delineated by the shape, or a holistic view of more than one of these dimensions. An alternative contrast condition which is a more defined categorical task would help to clarify these issues.

Experiment 2

In Experiment 1, we demonstrated that there are prototypical shifts using a serial study/test paradigm. To replicate the prototypical shifts and explore potential orienting task differences further, Experiment 2 was similar to Experiment 1 but replaced the preference judgment condition with an animacy judgment task. Judging the animacy (living or nonliving) status of a silhouette is a categorization task like the color judgment. Responding to animacy is a task which people—even infants—can perform easily (Mandler & Bauer, 1988).

The key information from the stimuli for the animacy task is the shape of the silhouette rather than the color. In preference judgments, it is possible that color plays a significant role in the judgments. Liking or disliking an item is a holistic judgment which would likely take into account the combination of the hue and shape. The separation of key information—hue for color judgments, shape for animacy judgments—for the two orienting tasks helps increase the difference between tasks. The minimum specificity alternative would suggest little to no difference between conditions: distinguishing between red and green takes little specificity of hue processing and distinguishing between living and nonliving shapes requires no specificity of hue processing. The directed attention alternative would predict higher accuracy for hue in the color judgment task, where hue is highlighted over shape. The results of this experiment can confirm the pattern found in Experiment 1 which is inconsistent with the representational shift hypothesis and lay groundwork for distinguishing between specificity and directed attention. Although this experiment alone cannot conclusively distinguish between the specificity and directed attention accounts, a difference between conditions would be more easily explained by the directed attention account.

Method

Forty undergraduate students were recruited as in the previous experiment. Participants answered the judgment questions “What color is this?” and “Is this a living thing?” in alternating blocks during the study phase. Otherwise, the stimuli and procedure were the same as in Experiment 1.

Results and Discussion

Twelve participants were excluded for low (<80%) color accuracy, low (<70%) animacy accuracy, or response patterns such as always pressing the same response key. Hence, 28 participants were included in the analyses reported, although the conclusions are the same based on data from all the participants. The included participants had 94.38% color accuracy and 88.84% animacy accuracy.

Participants showed systematic biased memory for hue in both the label and animacy conditions. The prototypical shifts observed in Experiment 1 were observed again. The color condition mean was 0.27 (standard deviation of 0.50) which is significantly more prototypical than the original hue value 0, $t(27) = 2.86$, $p < .01$, Cohen's $d = 0.54$, $BF_{10} = 5.54$. The animacy condition mean was 0.52 (standard deviation of 0.36) which is also significantly more prototypical than 0, $t(27) = 7.54$, $p < .001$, $d = 1.43$, $BF_{10} > 1,000$. Both distributions failed the Shapiro-Wilk test, both p values $> .05$. Comparing the proportion of accurate responses (Original column, Table 1), there was no difference by condition (paired, two-tailed $t(27) = 0.20$, $p > .05$, $d = 0.04$, $BF_{10} = 0.20$). Again, there was no evidence that any differences by condition are due to having more memory for hue/shape pairings in one condition over the other.

We again found color labeling does not uniquely cause response shifts, and further, it leads to weaker shifts than another kind of judgment. The animacy mean response typicality 0.52 was statistically different from the labeling mean response typicality 0.27, $t(27) = 3.12$, $p < .01$, $d = 0.59$, $BF_{10} = 9.53$, with animacy-judged hues shifting more strongly in a prototypical direction than color-judged hues. The BF between 3 and 10 indicates substantial support for a difference (Jeffreys, 1961). A chi-square analysis comparing the distribution of responses at each typicality level between the two conditions (see Table 1 for response rates) was also significant, $\chi^2(4, N = 1,120) = 10.916$, $p < .05$. These results point to different patterns of bias for the two encoding conditions with labeled hues being less biased toward the prototype than nonlabeled hues.

RTs followed the same pattern as in Experiment 1. Color judgment RTs were significantly faster (median: 444.71) than animacy judgment RTs (median: 528.93) according to a Wilcoxon signed-ranks test, $V = 0$, $p < .001$. Whereas animacy is a categorization task like color, the categorizations are at different hierarchical levels with color at a basic level and animacy at a superordinate level. The longer RTs for superordinate categorizations compared with basic categorizations is expected (Rosch, Mervis, Gray, Johnson, & Boyes-Braem, 1976). There was no statistical difference between the conditions for memory test RTs, $V = 142$, $p = .171$ (color median: 4,221.55, animacy median: 4,677.50).

Experiment 2 replicated the prototypical shift finding from Experiment 1 using an alternative orienting task. Responses on a memory test for hue systematically shift at this timescale. Again, both orienting tasks tested led to bias toward the prototype. Here the strength of shifts between conditions was significantly different with animacy judgments leading to more biased memory than color labeling. The prediction of the representational shift hypothesis, that overt labeling strengthens category prototype information during encoding, was again not supported. Still, this experiment does not conclusively distinguish between the directing attention and specificity hypotheses for the role of labeling. The condition difference here could be due to attention focusing on the color and, separately, the shape in the color and animacy judgment tasks, respectively. Alternatively, labeling color could set a minimum specificity of encoding that was greater than for the animacy condition for which the hue was irrelevant. In Experiment 3, we employed a nonlabel manipulation to further distinguish between alternative accounts.

Experiment 3

Why would labeled hues have a tendency for weaker shifts than in conditions without labeling? One possible explanation is that the labeled feature is processed to a minimum specificity for distinguishing between the alternatives. Since the labels used here (green or red; living or nonliving) are broad basic category labels, this could result in lower accuracy than if a more stringent subcategory criteria were applied. Another possibility is that labels direct attention to the appropriate features causing hue to be remembered with greater accuracy in the color labeling condition.

In Experiment 3, participants were informed of the upcoming memory test, including being shown a display of how narrow the differences between hue options are at test. By informing participants of the hue memory test prior to the orienting tasks, and showing them how highly specific the hue test would be, the knowledge of this test could define the focus of attention and the specificity of encoding during the study phase of the experiment. If the effect of labeling is actually an effect of attention to highlighted features, the higher accuracy found in the color condition in Experiment 2 should be achievable in the animacy condition if attention were directed to the hue feature by an alternate means. Therefore, both orienting conditions should produce weak shifts. If, instead, labeling sets the specificity of processing in the relevant dimension, alerting participants of a need for greater specificity than distinguishing between the label options would result in less shifting than the color labeling condition in Experiment 2.

Additionally, this manipulation of the instructions allows the question of labeling uniquely decreasing shifts to be examined. Eliminating the difference between the labeling and animacy conditions with an instructional (that is, a nonlabel-based) manipulation would suggest that labeling is not the only way to increase hue memory accuracy.

Method

Thirty-eight undergraduate students were recruited as in the previous experiments. The initial instructions, both presented verbally from the experimenter and written on screen, informed participants of the hue memory test that would follow the judgment task. The on-screen instructions displayed an example test question in order for participants to experience the small variation in hues displayed at test. Otherwise, the stimuli and procedure were the same as in Experiment 2 using the categorization and animacy conditions.

Results and Discussion

Eight participants were excluded using the same criteria as Experiments 1 and 2. Therefore, 30 participants were included in the analyses. Excluding participants did not change the pattern of results. The remaining participants had a color categorization accuracy of 91.42% and an animacy categorization accuracy of 88.00%.

We again found evidence of systematic prototypical shifts in the memory test response patterns for both the labeling and animacy conditions. The condition means were again significantly shifted toward the prototype (one sample, two-tailed t tests against original hue value of 0: color $M = 0.32$ ($SD = 0.43$), $t(29) = 4.07$, $p < .001$, $d = 0.74$, $BF_{10} = 88.57$; animacy $M = 0.26$ ($SD = 0.43$), $t(29) = 3.36$, $p < .01$, $d = 0.61$, $BF_{10} = 16.41$). Shapiro-Wilk tests did not

indicate nonnormal distributions, p values $> .05$. Comparing the original hue response rate by condition again yielded a null result (paired, two-tailed $t(29) = 1.21$, $p > .05$, $d = 0.22$, $BF_{10} = 0.38$).

Contrary to the possibility that the mitigating effect of labeling is unique to language, the conditions were not significantly different from each other (two-tailed, paired-sample t test: $t(29) = 0.63$, $p > .05$, $d = 0.12$, $BF_{10} = 0.23$; chi-square test: $\chi^2(4, N = 1,200) = 3.845$, $p > .05$; see Table 1 for response rates). The inverse Bayes factor (BF_{01}) is 4.27, which is considered to be substantial evidence for the null hypothesis that the conditions are not different. Forewarning the participants of the memory test affected participants' processing, seemingly in a similar way to the effect of labeling.

In Figure 2, it is clear that the mean response typicality for both conditions in this experiment mirror the color categorization condition mean response typicality in Experiment 2. This pattern was verified as statistically significant by conducting a 2×2 analysis of variance (ANOVA) with judgment condition within-subjects and instructional condition between-subjects (across experiments). There was an interaction of judgment and instructional conditions, $F(1, 56) = 6.752$, $p < .02$, $\eta^2 = 0.10$, $BF_{\text{Forward}} = 4.08$. The Bayesian analysis for the full model of both main effects and the interaction was inconclusive at $BF_{10} = 1.01$, but the interaction term BF_{Forward} of 4.08 indicates that just the interaction, not the full model, has substantial support against the null. This suggests that the instructional manipulation selectively and significantly changed the average typicality of the selected hue in the animacy judgment condition while not having much of an effect on the

color labeling condition. Although we acknowledge the possible limitations of cross-experiment statistical comparisons, we note that participants for the two experiments were drawn from the same pool and recruited in the same manner. This pattern is more consistent with the hypothesis that labeling directs attention to the relevant feature rather than setting a minimum specificity of processing.

RTs again followed the same pattern as Experiment 1 and 2 with significantly different RTs for the judgment responses (color median: 470.68, animacy median: 549.69, Wilcoxon signed-ranks test, $V = 6$, $p < .001$) and no statistical difference in RTs for the memory test responses (color median: 3,840.95 ms, animacy median: 3,920.65, Wilcoxon signed-ranks test, $V = 247$, $p = .777$). RT differences have been taken as a proxy for depth of processing, and the continued longer processing time for animacy would indicate more depth of processing. However, the combination of RT differences and a lack of difference in memory bias across conditions suggest there is not a linear relationship between length of processing and bias at least as is relevant to hue memory. Either processing time is not a good measure of depth of processing or depth of processing is not explanatory for the memory bias results.

Telling participants to attend to hue decreases hue memory response bias in this paradigm. The difference between orienting tasks disappears with both conditions leading to more accurate memory for hue. This lends support to the idea that rather than overt labeling selectively recruiting category knowledge in this paradigm, labeling acts as a cue to encode the hue feature of the exemplar with more precision. When making an animacy judgment the shape of the stimulus was most important for the task. Participants demonstrated they could, under suitable task instructions, attend to both hue and shape by maintaining a similar level of animacy judgment accuracy while also increasing their accuracy in hue memory.

Average Response Typicality by Judgment Condition

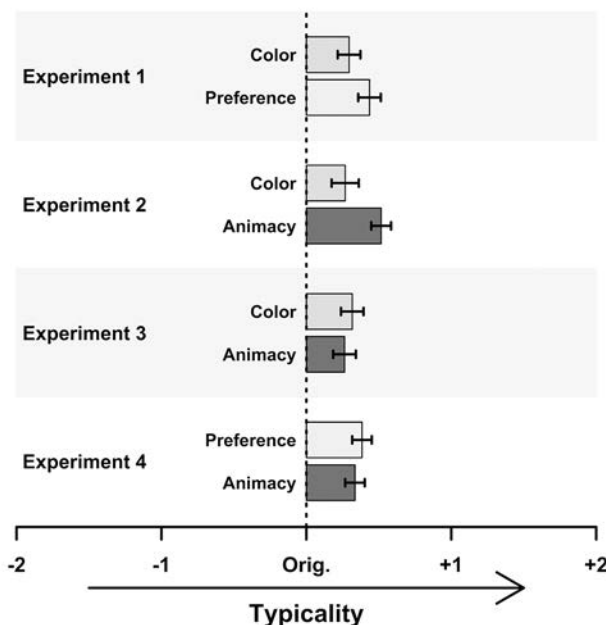


Figure 2. Mean typicality of response by condition. The dotted line represents the hues originally shown to the participant while the bars indicate the average typicality of the items with more typical hues to the right. This orientation visually displays the magnitude of the response shifts in hue typicality space. See the online article for the color version of this figure.

Experiment 4

Finally, we put the preference and animacy judgment conditions in the same experiment leaving out color labeling altogether. It has been suggested that language is a component of conceptual processing that can be up-regulated or down-regulated based on task demands (Lupyan, 2012). Specifically, engaging the language network for a task can globally recruit language-related processing to a task. Labeling color specifically enhanced hue memory response accuracy, decreasing shifting relative to nonlabeled stimuli in Experiments 1 and 2. Color labeling could have a specific effect of increased accuracy on labeling trials but could also have a global effect of priming color category knowledge driving the overall shifts observed in all experiments and conditions. If color labels are not overtly engaged, will systematic response shifts still occur? If categorization affects perception only when relevant linguistic categories are explicitly engaged, it would suggest a loose coupling of perception with conceptual effects led by language. If categorization affects perception regardless of linguistic engagement, it would suggest a stronger coupling of perception with conceptual information. Preference judgments and animacy judgments do not require participants to attend to the colors and their labels. The color and silhouette pairs were chosen to be arbitrary and not semantically indicative of each other. These conditions without a color labeling condition and without fore-

knowledge of a color memory test provide a context that does not require participants to attend to hue during the study portion of the experiment.

Method

Forty undergraduate students were recruited as in the previous experiments. Participants answered the judgment questions “Is this a living thing?” and “Do you like this?” omitting the color categorization condition. Otherwise, the stimuli and procedure were the same as in Experiment 1 and 2 with the hue memory test again a surprise following study.

Results and Discussion

Three participants had low animacy accuracy (<70%) and were excluded from the analysis leaving a sample size of 37. Excluding these participants did not change the pattern of results. Included participants had an animacy accuracy of 89.93%.

We again found evidence for systematic prototypical response shifts. The response means for each condition were both significantly different from the target response (two-tailed, one-sample *t* tests against original hue value of 0: preference $M = 0.38$ ($SD = 0.42$), $t(36) = 5.69$, $p < .001$, $d = 0.94$, $BF_{10} > 1,000$; animacy $M = 0.34$ ($SD = 0.43$), $t(36) = 4.97$, $p < .001$, $d = 0.82$), $BF_{10} > 1,000$. The response distributions for both conditions again were not significantly different from normal according to the Shapiro-Wilk normality test $p > .05$. There was no difference by condition for the objectively correct original hue response rate at test (paired, two-tailed $t(36) = 0.06$, $p > .05$, $d = 0.01$, $BF_{10} = 0.18$).

The mean typicality of response for the two conditions were not different from each other (paired sample, two-tailed *t* test: $t(36) = 0.68$, $p > .05$, $d = 0.11$, $BF_{10} = 0.22$; chi-square test: $\chi^2(4, N = 1120) = 1.294$, $p > .05$, see Table 1 for accuracy rates). Again, looking at the inverse BF, $BF_{01} = 4.55$, there is substantial evidence for the null hypothesis that there is no difference in magnitude of prototypical shifts between the two encoding conditions.

The RTs tell a similar story to the previous experiments. There was a significant difference between the more quickly responded to animacy condition (median: 523.36 ms) and the preference condition (median: 560.82 ms) according to the Wilcoxon signed-ranks test, $V = 518$, $p < .02$. This provides evidence that there is more depth of processing in preference judgments than animacy judgments. There was no significant difference in memory test RTs (animacy median: 4,099.75 ms, preference median: 3,813.05 ms, Wilcoxon signed-ranks test, $V = 338$, $p = .846$).

With no directing of attention to color either through instructions or through an overt labeling condition, participants did not have any indication that paying attention to color was important for a future task. The results of the memory test in this case continued to show systematic shifts of hue memory responses toward the prototype. Overt language was not required to obtain shifts relative to the category structure. The lack of an effect of condition is consistent with these encoding conditions equally not calling attention to hue. The shifts occurring regardless of color labeling indicates a strong influence of category knowledge in this task with overt priming of the labels not needed.

General Discussion

In these experiments we set out to investigate (a) if recognition memory for hue is systematically shifted toward the prototype, (b) how overt labeling of colors affects the directional shifts, and (c) if any effect of overt labeling on the shifts is unique. These questions were specifically investigated within a context of labels matching their referents to expand beyond known influences of language on memory when the language is contrastive.

Consistently, we have demonstrated that there is systematic bias to responses on a five-alternative forced-choice recognition memory test. Participants, on average, chose hues that were more prototypical of the basic color category than the hue that they had originally seen in all conditions in all four experiments. This result replicates the existence of response shifts we had previously demonstrated (Kelly & Heit, 2014). However, at the present timescale tested of a few minutes, the direction of the shifts was toward the prototype rather than away from the prototype as had been seen at the previous 500 and 5,000 ms timescales.

The present findings are novel evidence of prototypical shifts in hue memory test responses. By demonstrating the existence of prototypical response shifts at this timescale in all conditions, the results provide strong evidence of a systematic influence of categorical knowledge on hue recognition memory. Whether the prototypical response shifts are a demonstration of true shifts of encoded memory traces, or if they are due to response bias after memory is retrieved is still up for debate, a debate our additional results inform. Our further analyses and manipulations from experiment to experiment began to investigate the role of labeling in influencing the response shifts.

With regard to the role of language, we found that labeling does not produce stronger typical shifts than comparison conditions. Participants chose more typical hues at test when they had made preference or animacy judgments about a silhouette than when they had labeled the color category. The results do not support a special role of overt labels as top-down category prototypical attractors for memory traces prior to encoding as proposed by the representational shift hypothesis (Lupyan, 2008). Instead, labeling color categories had the opposite effect decreasing the magnitude of the shifts relative to other orienting tasks. This pattern was also not predicted by the specificity hypothesis, which would predict weak memory traces susceptible to the category information at response.

We followed up these findings by testing whether the label effect was driven by the third hypothesis, labels directing attention, while also more directly testing the specificity hypothesis. Our manipulations, informing participants of the upcoming memory test and showing participants an example test display, separated the minimum specificity needed by the task and the directing of attention to hue from the act of labeling (Experiment 3). Participants knew they needed to attend closely to hue with a high level of specificity and map that knowledge onto the shape cues. The manipulations eliminated the difference between conditions, and the response shifts in the two foreknowledge conditions of Experiment 3 were consistent with the color labeling condition of Experiment 2. Foreknowledge improved accuracy in the animacy condition but did not alter performance in the color labeling condition. This is consistent with the directing attention explanation of the label effect—participants when labeling color pay more

attention than they would when the shape is diagnostic for their labeling task, resulting in more accurate memory. If the specificity of processing idea were true, we would expect to have seen an improvement in hue recognition accuracy for the color labeling condition as well since the example display makes clear that much more specific knowledge of the hue is needed in the memory task than simply differentiation between *green* and *red*.

Additionally, Experiment 3 speaks against the idea that labeling has a unique effect on hue memory response shifts. While the previous experiments clearly did not support the idea of labels as prototypical attractors magnifying directional shifts, they could indicate labels uniquely decrease shifts. When given foreknowledge of the hue memory test, participants showed reduced shifting in the animacy condition matching the memory performance of the color labeling condition. This evidence speaks against the general hypothesis that labeling causes unique changes in the memory responses.

Finally, we tested whether explicit labeling during the task was required for the response shifts (Experiment 4). Lupyan (2012) suggested the influence of language can be up-regulated or down-regulated in the overall perceptual processing network. In terms of the present experimental design, color labeling in one condition would attune the system to the labeled categories and produce the overall category-relative shifting effect regardless of the trial condition. As mentioned earlier, hue memory has been shown to be both accurate (Pérez-Carpinell et al., 1998) and potentially distorted (Kay & Kempton, 1984; Winawer et al., 2007). We conducted a final experiment using the same paradigm comparing the two nonlabeling conditions with no overt mentions of the hue color labels and with the hue memory test a surprise. There continued to be an overall prototypical bias in hue memory indicating that explicit labeling during the task is not required to produce the effect.

Together, the results show a pervasive, systematic influence of category knowledge during a hue recognition memory test resulting in prototypical response shifts. Further, category-consistent labeling has an effect of decreasing bias, is not unique in producing the decreased effect, and is not required to produce category-relative response shifts. Labeling appears to efficiently direct attention to a particular feature of a stimulus relevant to that label, resulting in increased memory accuracy for that feature.

Shifts of What?

As can be seen in Table 1, in no condition across all four experiments did participants choose the original hue at even chance levels. Four lures and the original hue reduced the percentage each item would be chosen at random to 20%. In all conditions, the most typical lure, followed by the second most typical lure, was chosen most often. They were the only items chosen above chance.

As mentioned in the introduction, there are two possible interpretations of the response shifts. One possibility is memory traces themselves shift. While they do not shift only, or more strongly, when labeled, it is possible that all of the representations are shifting toward the prototype. In this interpretation, labeling the color category reduces the amount of shift of the memory trace by calling attention to the color feature of the silhouettes during initial encoding. Encoded memory would be influenced by the prototyp-

ical category information, with the labeled hues less influenced. The other possibility is memory is not shifting so much as there is a strong typical response bias that can overwhelm the memory trace. With no memory of the hue, or a memory not specific enough to distinguish between the hue options at test, participants tend to choose the hue that is most typical of the color category, which is also the central tendency of all hues of that color category presented during the experiment. Within this interpretation, labeled hues would have a stronger memory trace and would be less subject to the influence of the category during retrieval. In both cases, memory is improved rather than distorted by the act of labeling during encoding.

Limitations

Although our results of bias and mitigation of bias are notable within the scope of our investigation—hue space with basic color labels—we acknowledge that this does not necessarily translate to pervasive bias in perception and memory during everyday life. The current experiments constituted a laboratory-based investigation with some unnatural controls. Colors usually vary in brightness and saturation in addition to hue. They are normally seen under different lighting conditions. Objects will vary in dimensions other than shape and hue. Therefore, natural contexts and situations might not result in the same shifts we observed above.

For our paradigm, color space was a superior domain in comparison to object space due to its well-defined parameters and known average focal or prototypical coordinates for the basic color categories in English. However, color is one feature rather than a whole object and may behave differently. Additionally since color belongs to the class of categories that have cross-boundary categorical perception effects, shifts may happen more readily or may be more exaggerated than in other potential test spaces.

Labeling itself is a somewhat unnatural task. Not much time is spent overtly labeling the objects and features encountered in everyday life. More often labels are used as part of a broader discourse about the context speakers find themselves within. So while labeling has an effect of minimizing bias with the current paradigm one could imagine with a different paradigm language having less of an attention-directing effect. How language affects perceptual processing in natural perception of the world is not directly addressed by our controlled laboratory-based task.

Acknowledging these limitations, our aim with this set of experiments was to address the specific question of memory accuracy when labels are applied, continuing a recent line of inquiry (Blanco & Gureckis, 2013; Kelly & Heit, 2014; Lupyan, 2008; Richler et al., 2011). As discussed below and in the existing literature, the demonstrated shifts have theoretical implications that reach beyond this isolated laboratory task.

Prototypical Versus Atypical Shifts

Kelly and Heit (2014) showed atypical shifts for all items in all conditions using the same hues as were used in the present experiments. The experiments focused on testing the prediction of prototypical shifts occurring during initial processing (Lupyan, 2008). In Kelly and Heit (2014), the memory test was an immediate test after the encoding task at 500 ms or 5,000 ms with the encoding and test tasks interwoven. Participants were asked if the just previously dis-

played and judged hue, using color labeling and preference judgment conditions, was the same or different from the subsequent hue displayed while the participant responded. Kelly and Heit (2014) showed that shifts are a measurable phenomenon and that there was not evidence for the predicted prototypical shifts at this short timescale. It did not clarify whether shifts were exclusively atypical or what the role for labeling would be in a long-term memory task.

In the present experiments, we used a hybrid of the original paradigm from Lupyan (2008) and the paradigm we developed to detect shifts. We continued to use the domain of hue and kept the four lures rather than one, but reverted to the serial study block followed by test block structure. The present experiments were meant to test for prototypical shifts at a timescale similar to those used in studies finding a label effect and to investigate the role of language. Using this hybrid paradigm, prototypical shifts were evident.

We have not yet experimentally investigated the relationship between these two sets of results. It could be that one of the above parameters or task demands accounts for the difference in shift direction. We favor the longer time delay and the addition of an intervening task in the present experiments as the factors driving the differences. Kelly and Heit (2014) investigated immediate memory for hue. The present experiments required further processing of hue into long-term memory and necessitated retrieval. There are other instances of changing memory responses after time delays. The sleeper effect is a change in attitudes over time (Pratkanis, Greenwald, Leippe, & Baumgardner, 1988) with a persuasive message being followed by a discounting cue resulting in immediate low persuasion with higher persuasion over time. Pratkanis et al. (1988) suggested this is caused by two memories with differential decay combining to influence the result. Immediately the discounted message is not trusted but because the cue for the discounting information decays faster, the message becomes more trusted over time. Another example is a change in attributional judgments over time (Jacoby, Kelley, Brown, & Jasechko, 1989). Fame was attributed to names that were not the names of famous individuals after a delay. The names were initially presented on a list of nonfamous names. After a day, participants did not remember the source list for the names but the names were familiar, resulting in reports of the names being famous. This suggests that memory for a stimulus retains its context immediately but can lose context over time. This could be decoupling of an association or it could be applying categorical criteria at response that produce different categorizations than had been applied at encoding. The context of many similar hues may be task parameters that cause distancing or atypical estimation bias. The delayed recall may result in a loss of that context with participants choosing hues that are more familiar and more typical.

Broader Implications

Representational shifts have been the subject of debate in the past several years because their existence would have far-reaching implications for theories of categorization, memory, the influence of language on thought, and beyond. Previous work has discussed such implications from the point of view of the original representational shift hypothesis and the minimum specificity view. With the current experiments, it is clear neither previous theory accounts for our data. Below we discuss how the revealed memory response shifts, with weaker shifts due to labeling as a consequence of

directed attention, might affect our understanding of other cognitive processes and phenomena.

Category structure. Semantic memory can be conceived as being made up of memory traces or exemplars (Heit & Barsalou, 1996; Mack, Preston, & Love, 2013; Nosofsky, 1986; Palmeri & Nosofsky, 2001). Each trace is a building block that contributes to overall knowledge of the world. Small shifts during trace creation could represent a bias in the build-up of exemplars underpinning categories, particularly if the shifts are biased by the existing category structure. The newly formed shifted representations become part of the categories that influenced their shifting. These exemplars would proceed to influence the exemplar distribution making up the category. Rather than categories being an accumulation of raw statistical experience, the distortions in new exemplars would systematically affect the content of categories. The affected category distributions would proceed to influence shifts in future experiences creating a feedback loop. Even if the bias seen in the current experiments was exclusively introduced at response, memory of the response could override the previous memory or create another, as has been shown in memory reconsolidation research (Nader & Einarsson, 2010).

If new representations prove to be systematically shifted toward the prototype, the bias would be to maintain the preestablished categories through increased similarity at the expense of other possible knowledge organizations. If the representations were to shift systematically away from the prototype, as reported in Kelly and Heit (2014), the bias would undermine the current organization and support new ways of conceptualizing a semantic space. We have now shown both directions of shifts in different experimental conditions suggesting a dynamic, pervasive influence of past experience and existing categories on new representations. The context surrounding perception affects the direction of shifting and therefore the bias introduced to categories in memory. There appears to be a dynamic influence of the environment on the interplay between perception, memory, and language in real time.

Estimation bias. As mentioned in the introduction, CAM (Huttenlocher et al., 1991, 2000, 2007) appears to be consistent with the observed shifts. CAM is a model of estimating spatial location from memory which puts the introduction of bias during response. Items are encoded with some level of specificity as they are perceived. Then this exemplar information and the relevant category information combine while estimating and producing a response. The category information has more influence on less specific exemplars.

Applying a generalization of CAM to the current results, the hue exemplars given more attention would be encoded with more specificity in the bottom-up exemplar information, resulting in less ambiguity available to be affected by top-down category information. In the present results, when hues were featured at study either through labeling hues or through instructions to attend to hues, the exemplars were remembered with less bias, consistent with the bottom-up/top-down trade-off of CAM.

While the current results are broadly consistent with CAM, there are issues with applying the model directly. Specifically, CAM assumes that encodings are unbiased. Yet in Kelly and Heit (2014) we found atypical shifts near encoding. One possibility is encoding is not unbiased but can be influenced by tasks and other online cognitive processes. See Bae et al. (2015) for an application of CAM to color working memory at and near encoding.

To emphasize, we are not claiming that our results are exclusively accounted for by the CAM framework or that our results distinguish CAM from other competing models. We are seeking to illustrate that our results can be reasonably fit into an existing well-developed categorization framework. It is likely they can fit into others as well, such as a straightforward exemplar activation model, Minerva II (Hintzman, 1984), or a model that incorporates existing category information, the integration model (Heit, 1994).

Language and thought. Knowing a language with color category distinctions leads to categorical perception effects—faster, more accurate discrimination of items crossing a category boundary than equidifferent items, by some objective measure, that are within a category—while not having distinct categories in a language leads to an absence of categorical perception effects (Goldstone & Hendrickson, 2010; Roberson, Hanley, & Pak, 2009; Winawer et al., 2007). This is an example of language relativity, the idea that the language(s) an individual speaks affects the way in which they conceptualize the world (Carroll, 1956).

Labels guide and facilitate category formation (Lupyan et al., 2007; Gureckis & Goldstone, 2008). In the experiments reported above, all memory is shifting relative to the category prototypes. It is not a giant leap to suggest that because English has the categories red and green, hue memory moved relative to those categories. In Greek, which divides blue into two basic categories (Winawer et al., 2007), light blue and dark blue, or in Korean, which divides green into two basic categories (Roberson et al., 2009), bright green and dull green, it is likely that hue memory would move relative to those different prototypes and category boundaries for speakers of those languages. This is a question that could be tested empirically.

Presently, in addition to the overall shift toward the prototype and away from the boundary in line with linguistic category knowledge, there was a distinct role of overt labeling increasing memory accuracy. Online labeling appeared to direct attention to color, producing more accurate encoding. In this role of language, the things and situations a language has labels for, and perhaps habitually draws attention to, would be regularly encoded with more accurate detail than the things and situations that do not have labels.

Rather than being able to tell a simple story of language distorting reality toward the semantic knowledge encoded in the language, we have two opposing language effects. Globally, language helps guide categorization producing a memory distortion effect, yet also has a local, online effect of enhancing veracity, at least in the way language was used in the present task.

Conclusion

We have provided novel evidence for prototypical response shifts in memory for hue. These shifts can be seen and measured in a unidimensional testing space. Existing category knowledge influences and biases recognition memory responses. Category-consistent labeling as well as other means of drawing attention to a particular dimension of an object systematically enhances the accuracy of recognition memory for that dimension. The evidence does not suggest that perceptual processing of bottom-up information is hijacked by top-down prototypical knowledge activated by labels. Nor does the evidence suggest that labeling causes shallow percepts sufficient to determine the appropriateness of the label and no more. There are pervasive influences of existing knowledge on recognition. Language plays a role in modulating those influences. It is clear that even when

language is consistent with a percept at encoding, language influences how percepts are subsequently remembered.

References

- 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, 744–763. <http://dx.doi.org/10.1037/xge0000076>
- Blanco, N., & Gureckis, T. (2013). Does category labeling lead to forgetting? *Cognitive Processing*, 14, 73–79. <http://dx.doi.org/10.1007/s10339-012-0530-4>
- Brady, T. F., Konkle, T., Alvarez, G. A., & Oliva, A. (2008). Visual long-term memory has a massive storage capacity for object details. *Proceedings of the National Academy of Sciences*, 105, 14325–14329. <http://dx.doi.org/10.1073/pnas.0803390105>
- Carroll, J. B. (1956). *Language, thought, and reality; Selected writings of Benjamin Lee Whorf*. Cambridge, MA: MIT Press.
- Chomsky, N. (1975). *Reflections on language*. New York, NY: Pantheon Books.
- City University. (2002). *A Web-Based Colour Vision Test*. Retrieved from <http://www.city.ac.uk/avrc/colourtest.html>
- Fodor, J. (1983). *The modularity of mind*. Cambridge, MA: MIT Press.
- Goldstone, R. L., & Hendrickson, A. T. (2010). Categorical perception. *WIREs Cognitive Science*, 1, 69–78. <http://dx.doi.org/10.1002/wcs.26>
- Gureckis, T. M., & Goldstone, R. L. (2008). The effect of the internal structure of categories on perception. In B. C. Love, K. McRae, & V. M. Sloutsky (Eds.), *Proceedings of the 30th Annual Conference of the Cognitive Science Society* (pp. 843). Austin, TX: Cognitive Science Society.
- Hanley, J. R., & Roberson, D. (2011). Categorical perception effects reflect differences in typicality on within-category trials. *Psychonomic Bulletin & Review*, 18, 355–363. <http://dx.doi.org/10.3758/s13423-010-0043-z>
- Heit, E. (1992). Categorization using chains of examples. *Cognitive Psychology*, 24, 341–380. [http://dx.doi.org/10.1016/0010-0285\(92\)90011-P](http://dx.doi.org/10.1016/0010-0285(92)90011-P)
- Heit, E. (1994). Models of the effects of prior knowledge on category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 20, 1264–1282. <http://dx.doi.org/10.1037/0278-7393.20.6.1264>
- Heit, E., & Barsalou, L. W. (1996). The instantiation principle in natural categories. *Memory*, 4, 413–452. <http://dx.doi.org/10.1080/096582196388915>
- Hintzman, D. L. (1984). Minerva 2: A simulation model of human memory. *Behavior Research Methods, Instruments & Computers*, 16, 96–101. <http://dx.doi.org/10.3758/BF03202365>
- Huttenlocher, J., Hedges, L. V., & Duncan, S. (1991). Categories and particulars: Prototype effects in estimating spatial location. *Psychological Review*, 98, 352–376. <http://dx.doi.org/10.1037/0033-295X.98.3.352>
- Huttenlocher, J., Hedges, L. V., Lourenco, S. F., Crawford, L. E., & Corrigan, B. (2007). Estimating stimuli from contrasting categories: Truncation due to boundaries. *Journal of Experimental Psychology: General*, 136, 502–519. <http://dx.doi.org/10.1037/0096-3445.136.3.502>
- Huttenlocher, J., Hedges, L. V., & Vevea, J. L. (2000). Why do categories affect stimulus judgment? *Journal of Experimental Psychology: General*, 129, 220–241. <http://dx.doi.org/10.1037/0096-3445.129.2.220>
- Jacoby, L. L., Kelley, C., Brown, J., & Jasechko, J. (1989). Becoming famous overnight: Limits on the ability to avoid unconscious influences of the past. *Journal of Personality and Social Psychology*, 56, 326–338. <http://dx.doi.org/10.1037/0022-3514.56.3.326>
- Jeffreys, H. (1961). *Theory of probability*. Oxford, UK: Oxford University Press.
- Kay, P., & Kempton, W. (1984). What is the Sapir-Whorf hypothesis? *American Anthropologist*, 86, 65–79. <http://dx.doi.org/10.1525/aa.1984.86.1.02a00050>
- Kelly, L. J., & Heit, E. (2013). Representational shifts towards the proto-

- type in memory for hue. In M. Knauff, M. Pauen, N. Sebanz, & I. Wachsmuth (Eds.), *Proceeding of the 35th Annual Conference of the Cognitive Science Society*. Austin, TX: Cognitive Science Society.
- Kelly, L. J., & Heit, E. (2014). Representational shifts made visible: Movement away from the prototype in memory for hue. *Frontiers in Psychology*, 5, 796. <http://dx.doi.org/10.3389/fpsyg.2014.00796>
- Konkle, T., Brady, T. F., Alvarez, G. A., & Olivia, A. (2010). Scene memory is more detailed than you think: The role of categories in visual long-term memory. *Psychological Science*, 21, 1551–1556. <http://dx.doi.org/10.1177/0956797610385359>
- Loftus, E. F., & Palmer, J. C. (1974). Reconstruction of auto-mobile destruction: An example of the interaction between language and memory. *Journal of Verbal Learning and Verbal Behavior*, 13, 585–589. [http://dx.doi.org/10.1016/s0022-5371\(74\)80011-3](http://dx.doi.org/10.1016/s0022-5371(74)80011-3)
- Love, J., Selker, R., Marsman, M., Jamil, T., Dropmann, D., Verhagen, A. J., . . . Wagenmakers, E.-J. (2015). JASP (Version 0.7) [Computer software].
- Lupyan, G. (2008). From chair to “chair”: A representational shift account of object labeling effects on memory. *Journal of Experimental Psychology: General*, 137, 348–369. <http://dx.doi.org/10.1037/0096-3445.137.2.348>
- Lupyan, G. (2012). Linguistically modulated perception and cognition: The label-feedback hypothesis. *Frontiers in Psychology*, 3, 54. <http://dx.doi.org/10.3389/fpsyg.2012.00054>
- Lupyan, G., Rakison, D. H., & McClelland, J. L. (2007). Language is not just for talking: Redundant labels facilitate learning of novel categories. *Psychological Science*, 18, 1077–1083. <http://dx.doi.org/10.1111/j.1467-9280.2007.02028.x>
- Mack, M. L., Preston, A. R., & Love, B. C. (2013). Decoding the brain’s algorithm for categorization from its neural implementation. *Current Biology*, 23, 2023–2027. <http://dx.doi.org/10.1016/j.cub.2013.08.035>
- Mandler, J. M., & Bauer, P. J. (1988). The cradle of categorization: Is the basic level basic? *Cognitive Development*, 3, 247–264. [http://dx.doi.org/10.1016/0885-2014\(88\)90011-1](http://dx.doi.org/10.1016/0885-2014(88)90011-1)
- Mitterer, H., Horschig, J. M., Musseler, J., & Majid, A. (2009). The influence of memory on perception: It’s not what things look like, it’s what you call them. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 35, 1557–1562. <http://dx.doi.org/10.1037/a0017019>
- Nader, K., & Einarsson, E. Ö. (2010). Memory reconsolidation: An update. *Annals of the New York Academy of Sciences*, 1191, 27–41. <http://dx.doi.org/10.1111/j.1749-6632.2010.05443.x>
- Nosofsky, R. M. (1986). Attention, similarity, and the identification-categorization relationship. *Journal of Experimental Psychology: General*, 115, 39–57. <http://dx.doi.org/10.1037/0096-3445.115.1.39>
- Nosofsky, R. M. (2015). An exemplar-model account of feature inference from uncertain categorizations. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 41, 1929–1941. <http://dx.doi.org/10.1037/xlm0000120>
- Palmeri, T. J., & Nosofsky, R. M. (2001). Central tendencies, extreme points, and prototype enhancement effects in ill-defined perceptual categorization. *The Quarterly Journal of Experimental Psychology A: Human Experimental Psychology*, 54, 197–235. <http://dx.doi.org/10.1080/02724980042000084>
- Pérez-Carpinell, J., Baldoví, R., de Fez, M. D., & Castro, J. (1998). Color memory matching: Time effect and other factors. *Color Research and Application*, 23, 234–247. [http://dx.doi.org/10.1002/\(SICI\)1520-6378\(199808\)23:4<234::AID-COL8>3.0.CO;2-P](http://dx.doi.org/10.1002/(SICI)1520-6378(199808)23:4<234::AID-COL8>3.0.CO;2-P)
- Perfors, A., Tenenbaum, J. B., & Regier, T. (2011). The learnability of abstract syntactic principles. *Cognition*, 118, 306–338. <http://dx.doi.org/10.1016/j.cognition.2010.11.001>
- Persaud, K., & Hemmer, P. (2014). The influence of knowledge and expectations for color on episodic memory. In P. Bello, M. Guarini, M. McShane, & B. Scassellati (Eds.), *Proceedings of the Thirty-Sixth Annual Conference of the Cognitive Science Society* (pp. 1162–1167). Austin, TX: Cognitive Science Society.
- Pinker, S. (1994). *The language instinct: How the mind creates language*. New York, NY: Harper Perennial Modern Classics. <http://dx.doi.org/10.1037/e412952005-009>
- Pratkanis, A. R., Greenwald, A. G., Leippe, M. R., & Baumgardner, M. H. (1988). In search of reliable persuasion effects: III. The sleeper effect is dead. Long live the sleeper effect. *Journal of Personality and Social Psychology*, 54, 203–218. <http://dx.doi.org/10.1037/0022-3514.54.2.203>
- Reyna, V. F., & Brainerd, C. J. (1995). Fuzzy-trace theory: An interim synthesis. *Learning and Individual Differences*, 7, 1–75. [http://dx.doi.org/10.1016/1041-6080\(95\)90031-4](http://dx.doi.org/10.1016/1041-6080(95)90031-4)
- Richler, J. J., Gauthier, I., & Palmeri, T. J. (2011). Automaticity of basic-level categorization accounts for labeling effects in visual recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 37, 1579–1587. <http://dx.doi.org/10.1037/a0024347>
- Richler, J. J., Palmeri, T. J., & Gauthier, I. (2013). How does using object names influence visual recognition memory? *Journal of Memory and Language*, 68, 10–25. <http://dx.doi.org/10.1016/j.jml.2012.09.001>
- Roberson, D., & Hanley, J. R. (2010). Relatively speaking: An account of the relationship between language and thought in the color domain. In B. Malt & P. Wolff (Eds.), *Words and the world: How words capture the human experience* (pp. 183–198). New York, NY: Oxford University Press. <http://dx.doi.org/10.1093/acprof:oso/9780195311129.003.0010>
- Roberson, D., Hanley, J. R., & Pak, H. (2009). Thresholds for color discrimination in English and Korean speakers. *Cognition*, 112, 482–487. <http://dx.doi.org/10.1016/j.cognition.2009.06.008>
- Roediger, H. L., III, & McDermott, K. B. (1995). Creating false memories: Remembering words not presented in lists. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 21, 803–814. <http://dx.doi.org/10.1037/0278-7393.21.4.803>
- Rosch, E., Mervis, C. B., Gray, W. D., Johnson, D. M., & Boyes-Braem, P. (1976). Basic objects in natural categories. *Cognitive Psychology*, 8, 382–439. [http://dx.doi.org/10.1016/0010-0285\(76\)90013-X](http://dx.doi.org/10.1016/0010-0285(76)90013-X)
- Shepard, R. N. (1967). Recognition memory for words, sentences, and pictures. *Journal of Verbal Learning and Verbal Behavior*, 6, 156–163. [http://dx.doi.org/10.1016/s0022-5371\(67\)80067-7](http://dx.doi.org/10.1016/s0022-5371(67)80067-7)
- Sperling, G. (1960). The information available in brief visual presentations. *Psychological Monographs: General and Applied*, 74, 1–29. <http://dx.doi.org/10.1037/h0093759>
- Sturges, J., & Whitfield, T. W. A. (1995). Locating basic colours in the munsell space. *Color Research and Application*, 20, 364–376. <http://dx.doi.org/10.1002/col.5080200605>
- Tomasello, M. (2009). *Constructing a language*. Cambridge, MA: Harvard University Press.
- 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 of the United States of America*, 104, 7780–7785. <http://dx.doi.org/10.1073/pnas.0701644104>

(Appendices follow)

Appendix A

Stimuli Details

The stimuli for this set of experiments were complex. More detail on the specific items and hues is reported here. Additionally, to alleviate concerns that individual hues might be eliciting different categorizations, we reanalyze encoding data from the main experiments to detect if there are any differences in judgments by hue.

Item Randomization

Forty silhouette shapes were created with 20 living things and 20 nonliving things listed in Table A1. Eight stimulus lists were created rather than a full randomization. Five living things and five nonliving things were assigned to each of four sets.

Table A1
Item Randomizations

Set No.	Silhouette	Hue		Condition			
		1	2	A	B	C	D
Set 1	<i>Seahorse</i>	g1	r2	Color	Other	Other	Color
	<i>Butterfly</i>	g2	r3	Color	Other	Other	Color
	<i>Giraffe</i>	g1	r4	Color	Other	Other	Color
	<i>Moose</i>	g2	r2	Color	Other	Other	Color
	<i>Seal</i>	g2	r2	Color	Other	Other	Color
	Sunglasses	g3	r2	Color	Other	Other	Color
	Sailboat	g3	r3	Color	Other	Other	Color
	Car	g1	r4	Color	Other	Other	Color
	Watering can	g1	r3	Color	Other	Other	Color
	Hammer	g4	r1	Color	Other	Other	Color
Set 2	<i>Rat</i>	r4	g1	Color	Color	Other	Other
	<i>Fish</i>	r2	g4	Color	Color	Other	Other
	<i>Cat</i>	r3	g3	Color	Color	Other	Other
	<i>Parrot</i>	r2	g3	Color	Color	Other	Other
	<i>Bat</i>	r4	g3	Color	Color	Other	Other
	Scissors	r3	g4	Color	Color	Other	Other
	Shirt	r1	g4	Color	Color	Other	Other
	Airplane	r3	g1	Color	Color	Other	Other
	Sword	r3	g4	Color	Color	Other	Other
	Lamp	r2	g3	Color	Color	Other	Other
Set 3	<i>Snail</i>	g3	r4	Other	Color	Color	Other
	<i>Horse</i>	g2	r3	Other	Color	Color	Other
	<i>Cow</i>	g2	r4	Other	Color	Color	Other
	<i>Polar bear</i>	g2	r1	Other	Color	Color	Other
	<i>Rabbit</i>	g3	r1	Other	Color	Color	Other
	Genie lamp	g4	r3	Other	Color	Color	Other
	Key	g3	r2	Other	Color	Color	Other
	Cloud	g2	r4	Other	Color	Color	Other
	Pan	g4	r1	Other	Color	Color	Other
	Badge	g4	r1	Other	Color	Color	Other
Set 4	<i>Camel</i>	r4	g4	Other	Other	Color	Color
	<i>Penguin</i>	r1	g1	Other	Other	Color	Color
	<i>Kangaroo</i>	r1	g2	Other	Other	Color	Color
	<i>Rhino</i>	r3	g1	Other	Other	Color	Color
	<i>Gorilla</i>	r4	g3	Other	Other	Color	Color
	Balloon	r2	g2	Other	Other	Color	Color
	Anchor	r1	g4	Other	Other	Color	Color
	Trophy	r1	g1	Other	Other	Color	Color
	Satellite dish	r2	g1	Other	Other	Color	Color
	Office chair	r4	g2	Other	Other	Color	Color

Note. Living items for the animacy condition are italicized.

(Appendices continue)

Table A2
Hue Coordinates in LCH and Lab Color Spaces

Hue		L*	C	H	L*	a	b
g1--		42.842	42.909	135	43	-30	30
g1-	g2--	42.842	42.909	139	43	-32	28
g1	g2-	42.842	42.909	143	43	-34	26
g1+	g2	42.842	42.909	147	43	-36	23
g1++	g2+	42.842	42.909	151	43	-38	21
	g2++	42.842	42.909	155	43	-39	18
<i>Focal green</i>		<i>42.842</i>	<i>42.909</i>	<i>157.039</i>	43	-40	17
g3++		42.842	42.909	159	43	-40	15
g3+	g4++	42.842	42.909	163	43	-41	13
g3	g4+	42.842	42.909	167	43	-42	10
g3-	g4	42.842	42.909	171	43	-42	8
g3--	g4-	42.842	42.909	175	43	-43	4
	g4--	42.842	42.909	179	43	-43	1
r1--		41.221	79.347	14	41	77	19
r1-	r2--	41.221	79.347	18	41	75	25
r1	r2-	41.221	79.347	22	41	73	30
r1+	r2	41.221	79.347	26	41	71	35
r1++	r2+	41.221	79.347	30	41	69	40
	r2++	41.221	79.347	34	41	66	45
<i>Focal red</i>		<i>41.221</i>	<i>79.347</i>	<i>36.143</i>	41	64	47
r3++		41.221	79.347	38	41	62	49
r3+	r4++	41.221	79.347	42	41	59	53
r3	r4+	41.221	79.347	46	41	55	57
r3-	r4	41.221	79.347	50	41	51	61
r3--	r4-	41.221	79.347	54	41	46	64
	r4--	41.221	79.347	58	41	42	67

Note. The bold hues were shown at study. The italicized focal hues are for reference; Participants did not see the focal hues.

Each silhouette shape was randomly assigned a green and a red target hue. Two of the sets were green for each participant and one of the green sets would be in the color categorization condition and the other would be in the preference (Experiment 1) or animacy condition (Experiments 2 and 3). For example, reading Table A1, in Condition 2C, Set 1 and Set 2 are red and green, respectively. They are both in the preference/animacy condition. In Experiment 4, the same lists were used with the preference condition using the “color” lists from the table and the animacy condition using the “other” lists.

Color Calculations

The hues were calculated in L*CH color space then translated into L*ab color space for input into Adobe Photoshop, as reported in Table A2. The device-independent colors were translated into RGB device-dependent color values via individual monitor color profiles created for three Dell UltraSharp U2410 monitors by an X-rite i1 Display Pro color calibrator to ensure color constancy across screens.

In L*CH color space, hue is represented by degrees of a circle. The target hues were chosen both clockwise and counterclockwise from the focal colors reported in Sturges and Whitfield (1995). The

saturation (C) and lightness (L) were held constant at the focal values.

Hue Accuracy Analysis

An assumption of our study is that participants could consider the displayed hues to be members of their intended category. We did a cross-experiment analysis of color naming accuracy to determine if any hues were less likely to be named accurately. Not every participant saw every hue in the naming condition—that is, in Table A2, the green Sets 2 and 3 are each missing one of the four green targets, a product of randomization. We limited the analysis to just those participants who saw the full range of hues in the naming condition. We also continued to exclude participants with low accuracy and/or response patterns indicating lack of engagement. Within this group, accuracy across all hues was 92.7% with individual hue accuracies ranging from 88.9% to 95.6%. Because Experiment 3 had different instructions before the naming task, which could potentially influence performance during the encoding task, we included instruction condition as a factor. In a two-way ANOVA on naming accuracy, with the eight target hues as

(Appendices continue)

Table A3
Response Rates by Condition and Hue Category

Judgment condition	Hue category	Proportion correct (<i>SD</i>)
Color	Green	.62 (.17)
	Red	.69 (.15)
Living	Green	.67 (.16)
	Red	.68 (.15)
Overall		.67 (.16)

within-subject factors and two instruction conditions as a between-subjects factor, there was no difference in naming accuracy by hue, $F(3.95, 169.71) = 0.71$, $p > .05$, $\eta^2 = 0.02$, $BF_{10} = 0.02$, using the Greenhouse-Geisser correction for

violating the sphericity assumption. Based on the inverse Bayes factor, $BF_{01} = 47.46$, there is very strong evidence that there are not differences by hue. There was no main effect of instruction condition nor was there an interaction effect (both $p > .05$). These results indicate that participants were able to reliably identify less typical hues as members of the intended category as well as they could identify the more typical hues.

Similarly, there was no significant differences in proportions of “like” responses by hue in participants seeing all hues in the preference condition in Experiments 1 and 4, $F(3.87, 116.09) = 1.54$, $p > .05$, $\eta^2 = 0.05$, $BF_{10} = 0.15$, again using the Greenhouse-Geisser correction. This result indicates that there were also no systematic biases in preference judgments for the different hues, for example, participants did not like the “redder” reds better (Table A3).

Appendix B

Control and Norming Experiments

Two additional experiments were conducted, a control experiment accessing memory for the color category of a silhouette at test, and a norming experiment confirming that our calculated typical and atypical foils were, in fact, more or less typical of the color category than their target hue.

Category Memory Test

In the experiments reported above, participants made judgments about colored silhouettes and were tested on their hue memory in fine detail choosing among five hues spaced apart 2° in L^*CH color space along the circular hue dimension, approximately just noticeable differences. Each shape was presented in the original hue seen by the participant along with that target’s foil hues. However, by giving participants the category information at test and forcing a response among these slightly different hues out of category errors were not possible. Therefore, there is no measure of memory for the color category itself. Perhaps participants had no memory of the color of the silhouette at all and were guessing completely, with a typicality bias. To investigate color and shape bindings in our materials, we conducted a post hoc control experiment examining color category memory for the silhouettes at test. We used the same study procedure as Experiment 2, replacing the five-alternative forced-choice recognition memory test with a cued category memory test.

Method

Participants. Sixty-nine undergraduate students were recruited from UC Merced’s participant pool using the same criteria as in Experiments 1–4.

Stimuli. The study portion of the experiment used the target hues as in Experiments 1–4. The test portion of the experiment used gray (L^*ab hue: 82,0,0) versions of the silhouettes.

Procedure. Participants underwent the instruction and study procedure of Experiment 2. They made color and animacy judgments, and they had no foreknowledge of a memory test. Immediately after study, there was a memory test block. Each trial consisted of a light gray version of one of the silhouettes. Participants indicated if the silhouette had been green or red earlier with a button push. All 40 silhouettes were tested, with the order randomized for each participant. The response keys were the same as the color-labeling task during study, which were counterbalanced between participants. After completing this task they continued on to the typicality task reported below.

Results and Discussion

Ten participants had color categorization accuracies below 80% and/or animacy categorization accuracy below 70%. These participants were excluded resulting in a sample of 59 participants.

Overall, responses were 67% correct. There was an overall significant difference between the percent correct and chance (one sample t test, $\mu = 0.5$, $t(58) = 12.06$, $p < .001$, $d = 1.57$, $BF_{10} > 1,000$). We used a 2×2 repeated-measures ANOVA with color (red, green) and condition (color, animacy) as within-subject factors to investigate whether these factors affect color category memory. There was a main effect of color, $F(1, 58) = 5.09$, $p < .05$, $\eta^2 = 0.08$, $BF_{\text{Forward}} = 1.55$, with silhouettes correctly remembered as red more often than correctly remembered as green, though the Bayesian analysis suggests the difference is only

(Appendices continue)

slightly more likely to occur if the effect is included in the model. There was no main effect of condition, $F(1, 58) = 1.43$, $p > .05$, $\eta^2 = 0.02$, $BF_{\text{Forward}} = 0.23$ nor an interaction effect, $F(1, 58) = 2.82$, $p > .05$, $\eta^2 = 0.05$, $BF_{\text{Forward}} = 1.01$. Encoding condition did not affect the memorability of the color category of a silhouette; in fact, the Bayesian analysis favors the null hypothesis with a 1:0.23 or an approximate 81% likelihood the condition factor will not improve the model.

Typicality Norming

Color display can be difficult to control. The focal, or the most prototypical, colors we used as the basis for determining all other hues and determining the direction of “toward the prototype” were device-dependent values translated into device-independent space as discussed in Appendix A. This transformation, along with issues of ambient lighting, could result in inaccurate focal hue coordinates. We calibrated the displays and only used experiment rooms with approximately equal light intensity. But, to address concerns that our final displayed hues may not follow the typicality gradient assumed, we conducted a norming experiment. Participants were shown the hues used in Experiments 1–4 and asked to rate their typicality. We then analyzed these to determine if hues we calculated to be less typical of the category than their target hue and hues we calculated to be more typical than their target hue are, in fact, less and more typical.

Method

Participants. The same 69 undergraduates who participated in the category memory test above completed this experimental task.

Stimuli. The 24 hues were calculated and calibrated as described in Appendix A. Rather than silhouettes, the hues were displayed on uniform squares. The hues were displayed on the same monitors and in the same experiment rooms as were used in Experiments 1–4.

Procedure. Each trial consisted of a hue square with the question: “How typical is this of the color red[green]?” above the

square, and the response instruction: “Please respond on a scale of 1 to 5. 1 – Very Typical, 3 – Medium Typical, 5 – Not Typical At All” below the square. Each hue was displayed twice and the hues were displayed in blocks by color category. Therefore there were two blocks, one red and one green, each consisting of 24 trials. The blocks were counterbalanced between participants. The 12 hues of a color category were randomized twice, with all hues being displayed once before a second randomized order was displayed in the same block. This task occurred after the category memory task, and was followed by the CITY color blindness test (City University, 2002).

Results and Discussion

One participant had a response pattern indicating confusion with the task instructions. This participant was excluded along with the exclusions from the category memory task above resulting in a sample size of 58.

In order to determine relative typicality for each set of five hues, one target and its four foils, a difference score was computed subtracting the average typicality of the two foils we label more typical from the average typicality of the two foils we label less typical, for each set of five hues used in the memory test. Positive difference scores were consistent with our assumed typicality structure. There was an overall average difference of 0.68, which is significantly above 0, $t(57) = 12.69$, $p < .001$, $d = 1.67$, $BF_{10} > 1,000$. Red hues had a larger difference with a mean difference of 1.04 which was different from 0, $t(57) = 13.00$, $p < .001$, $d = 1.71$, $BF_{10} > 1,000$, while green hues had a mean difference of 0.32 which was also significantly different from 0, $t(57) = 4.87$, $p < .001$, $d = 0.64$, $BF_{10} > 1,000$.

These results indicate that even if our displayed hues were not perfectly representative of the calculated device-independent hue, they were on an appropriate atypical- to typical-gradient to meet the needs of our experimental design.

Received June 30, 2015

Revision received August 23, 2016

Accepted October 1, 2016 ■