Cognition 196 (2020) 104135

Contents lists available at ScienceDirect

Cognition

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

Finding categories through words: More nameable features improve category learning

Martin Zettersten*, Gary Lupyan

Psychology Department, University of Wisconsin-Madison, 1202 W Johnson Street, Madison, WI 53706, USA

ARTICLE INFO

Keywords:
Categorization
Category learning
Language
Nameability
Hypothesis-testing
Rule learning

ABSTRACT

What are the cognitive consequences of having a name for something? Having a word for a feature makes it easier to communicate about a set of exemplars belonging to the same category (e.g., "the red things"). But might it also make it easier to learn the category itself? Here, we provide evidence that the ease of learning category distinctions based on simple visual features is predicted from the ease of naming those features. Across seven experiments, participants learned categories composed of colors or shapes that were either easy or more difficult to name in English. Holding the category structure constant, when the underlying features of the category were easy to name, participants were faster and more accurate in learning the novel category. These results suggest that compact verbal labels may facilitate hypothesis formation during learning: it is easier to pose the hypothesis "it is about redness" than "it is about that pinkish-purplish color". Our results have consequences for understanding how developmental and cross-linguistic differences in a language's vocabulary affect category learning and conceptual development.

1. Introduction

What makes some categories difficult to learn and others easy? Some factors that contribute to learning difficulty are straightforward: it is harder to learn the difference between the letters 'b' and 'd' than between 'b' and 'm' because the former letters are more perceptually confusable (both in terms of their visual structure and in terms of their phonology). Other studied factors include the complexity of the rule that specifies category membership, the shape of the decision boundary, and the category covariance structure (Alfonso-Reese, Ashby, & Brainard, 2002; Feldman, 2000; Hahn, Chater, & Richardson, 2003; Nosofsky, Gluck, Palmeri, McKinley, & Glauthier, 1994; Shepard, Hovland, & Jenkins, 1961). The more complex the rule, the harder it will be to learn or induce. This idea is well expressed by Feldman: "Some concepts, by their nature, reduce to a very simple rule that covers all their members (like *red things*)." (2003, p. 231).

Characterizing categorization complexity in this way, however, leads to a puzzle. Consider the two categories shown in Fig. 1A. What is the intrinsic structure of these categories? Representing the category boundary as spanning two independent dimensions (horizontal displacement and vertical displacement, Fig. 1B) would suggest that learning this distinction involves integrating along two dimensions. But for someone familiar with English letters, there exists a one-dimensional, easily verbalized alternative: T-like vs. L-like (1C). This example highlights a general problem: *any* attempt to characterize a category structure requires first specifying a set of features (i.e., a vocabulary). But where do these features come from?

Often, models of categorization skirt this question by assuming that the features exist a priori. In models of categorization such as COVIS (Ashby, Alfonso-Reese, Turken, & Waldron, 1998), the features are often taken to correspond to perceptual primitives. For example, two widely used dimensions in category learning tasks—orientation and spatial frequency—map onto the main dimensions represented in primary visual cortex. Alternatively, the features are sometimes explicitly provided to the learner or the model simulating the learner (Alfonso-Reese et al., 2002; Love, Medin, & Gureckis, 2004). This (often tacit) assumption is common to otherwise very different models of categorization wherein the learner's task is viewed as learning how to appropriately weigh pre-existing features to correctly categorize the stimuli at hand (Carpenter, Just, & Shell, 1990; Kruschke, 1992; Nosofsky & Palmeri, 1996; Shiffrin & Steyvers, 1997; see

* Corresponding author.
E-mail address: zettersten@wisc.edu (M. Zettersten).

https://doi.org/10.1016/j.cognition.2019.104135
Received 14 March 2018; Received in revised form 7 November 2019; Accepted 11 November 2019

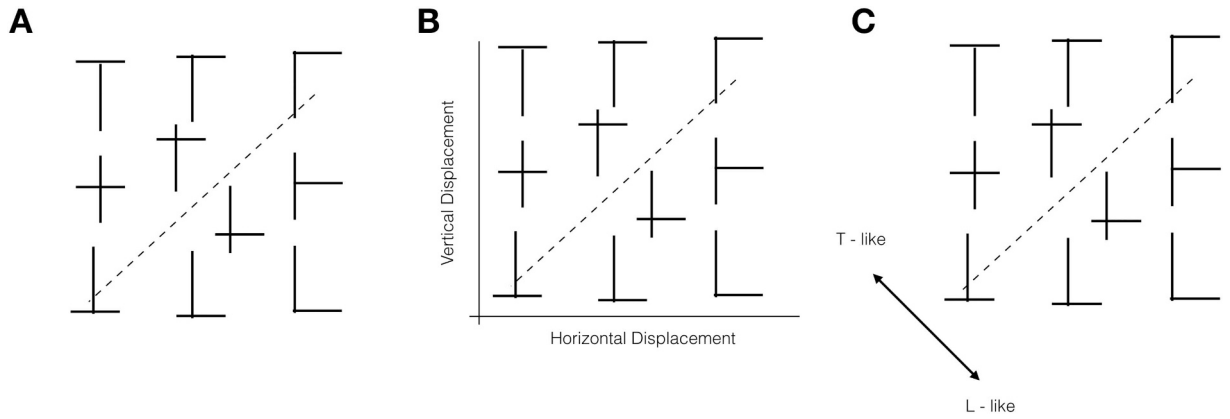

Fig. 1. Two-dimensional (B) and one-dimensional (C) representations of (A).

Goldstone, 2000 for discussion).[1]

The hypothesis we test in the present work is that one source of the features (or dimensions) we use in categorization are the words of natural language. On this view, what makes some categories simple is that they utilize features that have been acquired in the course of learning the vocabulary of our language. These lexically derived features are then flexibly deployed when faced with learning novel categories. For example, in encountering the stimulus space of Fig. 1, the features used in categorization would not simply correspond to horizontal and vertical lines, but the letters "T" and "L", thereby collapsing the otherwise two-dimensional space to a one-dimensional space in which the categorical distinction is something like "T-like" vs. "L-like". In the next section, we review empirical literature that supports this hypothesis.

1.1. The role of natural language in categorization

The idea that words play a causal role in categorization is, of course, prominently featured in the writings of Whorf (1956), but it is hardly limited to literature on linguistic relativity. In their seminal paper on category learning, Shepard et al. (1961) commented on the high correlation between learnability of categories varying on logical complexity (simple features, disjunctions, etc.) and the simplicity of the rules that could be verbally stated by subjects, suggesting that were it not for language, the formation of such rules may not be possible (see also Kurtz, Levering, Stanton, Romero, & Morris, 2013). Although numerous studies have investigated the relationship between formal complexity and learnability (Alfonso-Reese et al., 2002; Carpenter et al., 1990; Chater, 1999; Feldman, 2000, 2006; Vigo, 2006), the observation about the possible *causal* role of language, although sometimes acknowledged as a logical possibility (Ashby et al., 1998, p. 446; cf. Couchman, Coutinho, & Smith, 2010), has been largely overlooked (though see e.g., Fotiadis & Protopapas, 2014).

One reason for this oversight is that research on both language learning and category learning has often been conducted within a tradition of viewing words as reflecting the natural joints of the world (Lupyan, 2016). In a statement making this position explicit, Li and Gleitman (2002, p. 266) write, "…linguistic categories and structures are more-or-less straightforward mappings from a preexisting conceptual space, programmed into our biological nature: humans invent words that label their concepts". On this view, the fact that English has the words "triangle", "red", and "five" is a simple consequence of the structure of the world: the words correspond to a priori categories that objectively exist. If it were true that names simply reflect pre-existing categories, then the idea that learning words has a strong causal influence on categorization would seem to hold little merit.

But although some words indeed reflect natural kinds, the categories picked out by most words are neither obvious nor inevitable (Goddard & Wierzbicka, 2014). If words truly mirrored the structure of nature, one would expect substantial agreement in lexical systems across the world. It is not surprising to find that a language spoken in a culture lacking cars does not have the words "sedan" and "hatchback" as part of its core vocabulary. But substantially different patterns of naming are also found across much more fundamental domains including spatial relations (Haun, Rapold, Janzen, & Levinson, 2011; Levinson & Wilkins, 2006), time (Boroditsky, Fuhrman, & McCormick, 2011; Boroditsky & Gaby, 2010), common actions (Majid, Bowerman, van Staden, & Boster, 2007; Slobin, Ibarretxe-Antunanõ, Kopecka, & Majid, 2014), kin relations (Kemp & Regier, 2012; Murdock, 1970), body parts (Enfield, Majid, & van Staden, 2006), basic shapes (Luria, 1976; Roberson, Davidoff, & Shapiro, 2002), and colors (Gibson et al., 2017; Majid et al., 2018; Roberson, Davidoff, Davies, & Shapiro, 2005; Steels & Belpaeme, 2005).

Such linguistic diversity challenges "the prevailing assumption [that] many important concepts can be easily identified because they are revealed by words" (Malt et al., 2015, p. 292). Consequently, if categories like "red things" or "triangles" are easy for people to induce, we cannot assume that it is because they are, by their nature, simple. Rather, it may be that they are *made* simple by the ability of human learners to use verbal labels to reduce the dimensionality of the stimulus space. Language may be particularly important for abstract and rule-based categories that require learners to generalize across perceptually dissimilar members. These types of categories are often difficult for young children to learn (Minda, Desroches, & Church, 2008; Nazzi & Gopnik, 2001; Rabi & Minda, 2014) and success is often linked to learning relevant labels (Christie & Gentner, 2014). Words, on this view, not only carve nature at its joints, but also carve joints into nature.

Showing that language plays a causal role in categorization requires showing that linguistic manipulations affect people's categorization performance. Some of our previous work provides evidence in support of this general hypothesis. For example, teaching labels for novel categories facilitates category learning. Controlling for stimulus familiarity and overt categorization experience, category learning can be

[1] One alternative to assuming pre-existing features is to study how context can shift the types of features learners will search for in a category learning task (Wisniewski & Medin, 1994) or by having people learn the features from perceptual and categorization experience de novo (Kellman & Garrigan, 2009; Schyns, Goldstone, & Thibaut, 1998). These proposals have been discussed most in cases where people unitize previously separable features over the course of extended perceptual learning. Importantly, this process is different from the much more rapid category learning we study here. However, the mechanism by which such unitization occurs may be quite similar to what happens when we first learn a word, e.g., in learning "dog" we "chunk" together a variety of previously separable feature values that correspond to dogs.

boosted substantially by providing people with novel names for difficult-to-verbalize categories (Lupyan & Casasanto, 2015; Lupyan, Rakison, & McClelland, 2007). Even after a category is well-learned, hearing its name can make features denoted by the name more perceptually salient (Lupyan, 2008; Lupyan & Spivey, 2010; Lupyan & Ward, 2013). If language is causally involved in categorization, we can also predict that perturbing language may perturb categorization. Indeed, the ability to induce simple rules is often compromised in adults with aphasia, being specifically linked to impairments in naming (Hjelmquist, 1989; Koemeda-Lutz, Cohen, & Meier, 1987; Lupyan & Mirman, 2013). In healthy adults, learning categories with more easily verbalizable membership criteria is disrupted by verbal interference (Minda et al., 2008; Waldron & Ashby, 2001). For example, verbal interference impaired the performance of college students on selecting which of three objects was the odd one out based on a specific, easily verbalizable dimension (Lupyan, 2009)—a disruption that mirrored a pattern shown by an individual with a severe naming impairment (Davidoff & Roberson, 2004). Additional evidence comes from a finding that impairing naming using noninvasive brain stimulation (tDCS) can, under some circumstances, impact categorization performance (Mirman, Thompson-Schill, Lupyan, & Hamilton, 2012; Perry & Lupyan, 2014).

1.2. The present studies

While the work summarized above provides convergent evidence for a link between language and categorization, the variety of methods used make it difficult to make strong inferences about the causal impact of lexicalization on category learning. Here, we test the specific hypothesis that the ease of learning certain category distinctions is predictable from the ease of naming the constituent features. In testing this hypothesis, we vary nameability while holding constant—to the extent possible—other factors that are expected to influence categorization difficulty, including logical complexity and perceptual discriminability (Table 1). To the extent that natural language vocabulary provides a set of candidate features (or priors) that learners consider (Lupyan & Clark, 2015), more nameable categories will be easier to learn. For example, it may be easier to form a rule that Category A has “red things” as compared to Category A has “greenish-yellowish things”.

2. Experiment 1A: learning categories of more and less nameable color features

Experiment 1A tested whether rule-based categories are easier to learn when the features composing the rule are more vs. less nameable. Participants were shown circles composed of three colors and had to learn which circles corresponded to category A or B. In one condition, the colors were easy to name. In the other, they were more difficult to name. In both cases, one of the colors was 100% predictive of category membership, making it possible to perform perfectly by learning a simple rule (e.g., “circles with blue belong to category A”). We hypothesized that categories would be easier to learn when the color features were easier to name, and therefore could be more readily formulated as a hypothesis about category membership.

2.1. Method

2.1.1. Participants

We recruited 50 participants (19 female; all native speakers of English) through Amazon Mechanical Turk (mean age: 32.0 years; range: 21–61 years). Participants were randomly assigned to the High Nameability Condition (n = 25) or to the Low Nameability Condition (n = 25) and were paid $0.60 for completing the task (average completion time: 4.0 min, SD = 1.7).

Table 1
Overview of key design features, sample sizes, and effect sizes across experiments.

Experiment	Feature	Design	Stimulus Set	Basis for matching high vs. low nameability stimuli	N (# excluded participants)	Effect size
1A	Colors	Between-subjects	Color Set 1; critical contrast in high nameability condition: blue vs. orange	Within-prototype discriminability (ΔE2000)	50 (0 exclusions)	d = 1.01 95% CI = [0.41,1.61]
1B	Colors	Between-subjects	Color Set 1; critical contrast in high nameability condition: red vs. brown	Within-prototype discriminability (ΔE2000); critical feature discriminability	50 (0 exclusions)	d = 0.62 95% CI = [0.04, 1.20]
2A	Colors	Within-subjects	Color Set 1; critical contrast in high nameability condition: red vs. brown	Within-prototype discriminability (ΔE2000); critical feature discriminability	39 (1 exclusion)	d_z = 0.95 95% CI = [0.57, 1.33]
2B	Colors	Within-subjects	Color Set 2	All pairwise color comparisons (ΔE2000); overall behavioral discriminability matched based on same/different RTs.	39 (1 exclusion)	d_z = 0.93 95% CI = [0.55, 1.30]
3A	Shapes	Between-subjects	Vanderplas & Garvin shapes; 3-feature exemplars	Number of points and edges; average shape skeleton description length	48 (0 exclusions)	d = 1.12 95% CI = [0.49,1.74]
3B	Shapes	Between-subjects	Vanderplas & Garvin shapes; 2-feature exemplars	average shape skeleton description length; overall behavioral discriminability matched based on same/different RTs.	120 (0 exclusions)	d = 0.93 95% CI = [0.55, 1.31]
4	Shapes	Between-subjects	Tangram shapes	Same shapes rotated	119 (0 exclusions)	d = 0.46 95% CI = [0.09, 0.83]

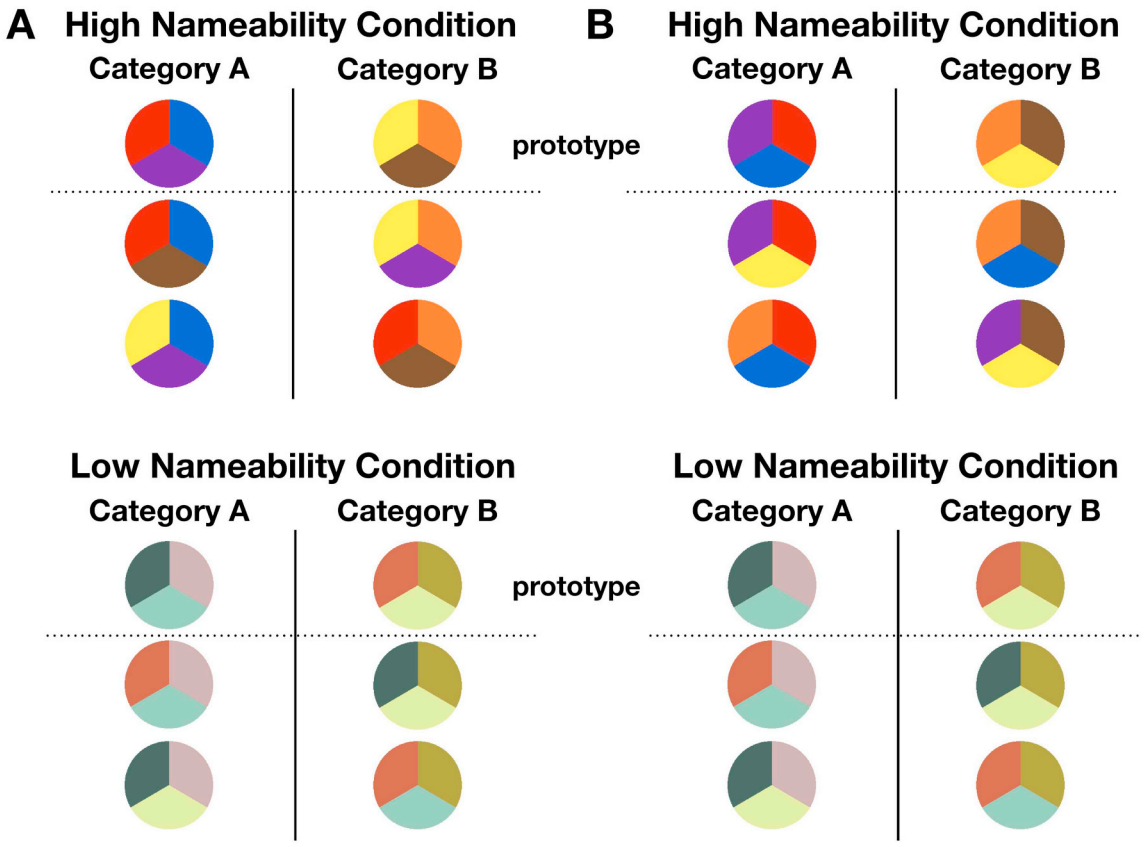

Fig. 2. Stimuli in Experiments 1A-B.

2.1.2. Stimuli

The exemplars were circles ("color wheels") composed of 3 different colors (see Fig. 2). Following the design of Couchman et al. (2010), one of the colors was perfectly predictive of category membership. The other two were correlated at 75% with category membership during training. Color pairs were tied to specific locations, e.g. the colors in the bottom slice of the circle were either "brown" RGB = (120, 80, 40) or "purple" RGB = (130, 30, 180). The position of the predictive color was always the top right slice of the circle. The stimulus composed of the three colors that occurred most frequently with each category was termed the "prototype". The two other stimuli in each category differed from the prototype with respect to one of the two 75% predictive colors (stimuli for all experiments can be downloaded from https://osf.io/fmhku/).

The critical manipulation involved the nameability (Guest & Van Laar, 2002) of the colors comprising each color wheel exemplar. To assess nameability, we used the results of a large-scale online study in which people were asked to name colors (Munroe, 2010). After removing likely spam and non-English responses, we were left with 2,947,648 naming trials from 134,727 participants. Each participant named between 1 and 425 colors (M = 22). Despite the amount of data, this is still a sparse sampling of the 16.7 million colors displayed by a typical computer screen, and so we rounded each color to the nearest 10 RGB value and restricted our search to the 120 colors named by ≥100 individuals with highest and lowest nameability, calculated based on the consistency with which people used the modal label for the color (see below).

There is no single agreed-upon way to quantify nameability. We considered two common approaches: modal agreement A (Brodeur, Dionne-Dostie, Montreuil, & Lepage, 2010; Brodeur, Guérard, & Bouras, 2014) and Simpson's diversity index D (Majid et al., 2018; Simpson, 1949).

Modal agreement is a frequently used measure of nameability that captures the degree to which participants converge on a common label for a stimulus. We computed modal agreement for a given item as the proportion of responses on which the most frequent unique response occurred. For example, if a given item was named "red" six times and "pink" four times, the modal agreement would be calculated as $A = 0.6$.

Simpson's diversity index was originally developed in ecology to measure species diversity in a way that accounted for both the number of types of species and the frequency of those types (Simpson, 1949). This measure has proven useful in assessing naming diversity in a manner that accounts for both type and frequency of labels for a stimulus (Majid et al., 2018; Majid & Burenhult, 2014). For instance, compare the hypothetical pattern of responses to a stimulus described above (response pattern 1: six "red" responses; four "pink" responses) to a case where there are a number of different alternative responses in addition to "red" (response pattern 2: six "red" responses; one response of "pink", "ruby", "carmine", "crimson" each). Modal agreement would be identical for these two response patterns ($A = 0.6$) despite the first case having only two unique responses and the second case having five. Simpson's diversity index takes into account the frequencies of each response type. Formally, for a given stimulus, if speakers produce N description tokens, including R unique description types from 1 to R, each with frequencies of n_1 to n_R, then Simpson's diversity index D is computed as

$$D = \frac{\sum_{i=1}^{R} n_i(n_i - 1)}{N(N-1)}$$

This measure ranges from 1 - indicating high nameability (all respondents gave the same response type, i.e. $i = 1$ and $n_i = N$) - to 0 - indicating low nameability (all respondents gave unique response types, i.e. $n_i = 1$ for all i). In the example from above, hypothetical

Table 2
Overview over the color feature set in Experiments 1A, 1B, and 2A.

RGB	Color	Modal name	Nameability	Modal agreement	Simpson's diversity	Discriminability RT	Role Exp 1A	Role Exp 1B & 2A
(30, 90, 210)		Blue	High	80.3%	0.671	557 ms	100% predictive	75% predictive
(250, 120, 30)		Orange	High	85.1%	0.733	575 ms	100% predictive	75% predictive
(220, 20, 0)		Red	High	82.7%	0.697	587 ms	75% predictive	100% predictive
(250, 240, 0)		Yellow	High	81.7%	0.664	568 ms	75% predictive	75% predictive
(120, 80, 40)		Brown	High	81.8%	0.648	577 ms	75% predictive	100% predictive
(130, 30, 180)		Purple	High	82.1%	0.672	578 ms	75% predictive	75% predictive
(170,160,40)		Mustard	Low	6.9%	0.056	587 ms	100% predictive	100% predictive
(200, 170, 170)		Grey	Low	6.8%	0.054	582 ms	100% predictive	100% predictive
(200, 100, 70)		Brown	Low	8.7%	0.051	554 ms	75% predictive	75% predictive
(70, 100, 90)		Grey green	Low	9.8%	0.128	575 ms	75% predictive	75% predictive
(220, 240, 150)		Pale green	Low	5.3%	0.079	579 ms	75% predictive	75% predictive
(150, 200, 180)		Green	Low	6.0%	0.084	580 ms	75% predictive	75% predictive

Note. See the Method section of Experiment 2A for details on the discriminability RT column.

response pattern 1 has a higher diversity index ($D = 0.47$) than hypothetical response pattern 2 ($D = 0.33$). Modal agreement and Simpson's diversity index are typically highly correlated, but Simpson's diversity index is particularly useful at differentiating nameability when there are a number of different response types (as is the case with the color and shape features used in the present studies). We thus use Simpson's diversity as our main measure of nameability in our analyses.

Notice that both measures rely on group naming behavior to make inferences about the ease with which an individual can name a stimulus. It need not be the case that low group name agreement corresponds to low individual nameability. Low group name agreement can occur if individuals have different, but nonetheless easily accessible names for the same entity. In practice, however, with the exception of stable idiolects wherein different individuals have different, but strongly preferred names for the same objects (e.g., "handbag" vs. "purse"), name agreement derived from group data turns out to strongly predict individual ease of naming (Balota et al., 2007; E. Bates et al., 2003; Brown & Lenneberg, 1954; Liu, Hao, Li, & Shu, 2011; Rossion & Pourtois, 2004; Székely et al., 2003). That is, if people don't agree on what to call something, it generally means that that something is also difficult for individuals to name.

Our goal was to select colors with high nameability and colors with low nameability for the two sets of prototype stimuli such that the three color features of each prototype stimulus had approximately equivalent pairwise CIE-LAB distances as quantified by ΔE2000 (Sharma, Wu, & Daa, 2005), to exclude the possibility that an advantage for highly nameable colors could be explained by their greater discriminability. To find a set of prototype stimuli with similar between-color discriminability, we randomly sampled sets of three colors (either all high nameability or all low nameability colors) from the remaining possible color combinations with the following constraints: each of the three colors must be clearly discriminable from the remaining two colors (ΔE2000 > 20) and the average ΔE2000 distance betweenhe color features comprising each prototype stimulus must lie between 35 and 45. Table 2 shows the resulting high and low nameability colors (Color Set 1) and the role each color served in the category structure used in each study (see also Fig. 2). The average within-prototype feature ΔE2000 discriminability was similar for high nameability colors (M = 39.7, SD = 11.6) and for low nameability colors (M = 36.5, SD = 9.5), $t(10) = 0.52$, $p = .61$. Note that the modal names for low nameability colors are sometimes frequent color terms that are poor labels for individual colors (e.g., "grey" for rgb(200,170,170)). This reflects the low agreement among participants in labeling these colors - individual responses were often more descriptive, but idiosyncratic (e.g., "slightly neutral lavender" or "light dusty rose" for rgb (200,170,170)). Modal responses for some low nameability items were two-word responses (e.g., "grey green" and "pale green"). See experiment 2A for more detailed information on the discriminability of the selected color features, including behavioral norming data.[2]

2.1.3. Design & procedure

2.1.3.1. Stimulus presentation. The stimuli were presented using a web-based task created using the jsPsych library (de Leeuw, 2014) and custom scripts for the category learning task.

2.1.3.2. Training design. The participants were instructed to place the color wheels into one of two categories by dragging the color wheel into either box "A" or box "B" (Fig. 3). Participants completed 24 training trials split into 3 blocks. On each block, participants sorted the prototype exemplar (the top image in Fig. 2A) of each category twice, and the remaining two exemplars of each category once. The order of the stimuli was randomized within each block. Participants were instructed to respond as quickly and accurately as possible and received immediate feedback on whether their choice was correct or incorrect. Trials were repeated after an incorrect response. Box locations (left/right) were counterbalanced across participants. At the end of the categorization task, we also asked participants how they decided whether to sort images into category A or category B and asked whether they used a particular strategy (see Section S2 in the Supplementary materials for further details and results).

2.2. Results

All data and R scripts of the analyses are openly available on OSF (https://osf.io/fmhku/). We fit a logistic mixed-effects model

[2] Accurate calculation of ΔE2000 requires a device-independent color space. Participants naming the colors viewed them on their (generally uncalibrated) monitors, making our ΔE estimates somewhat less reliable than estimates under more controlled viewing conditions. Note that in the lab-conducted experiments 2A and 2B, lighting conditions were constant.

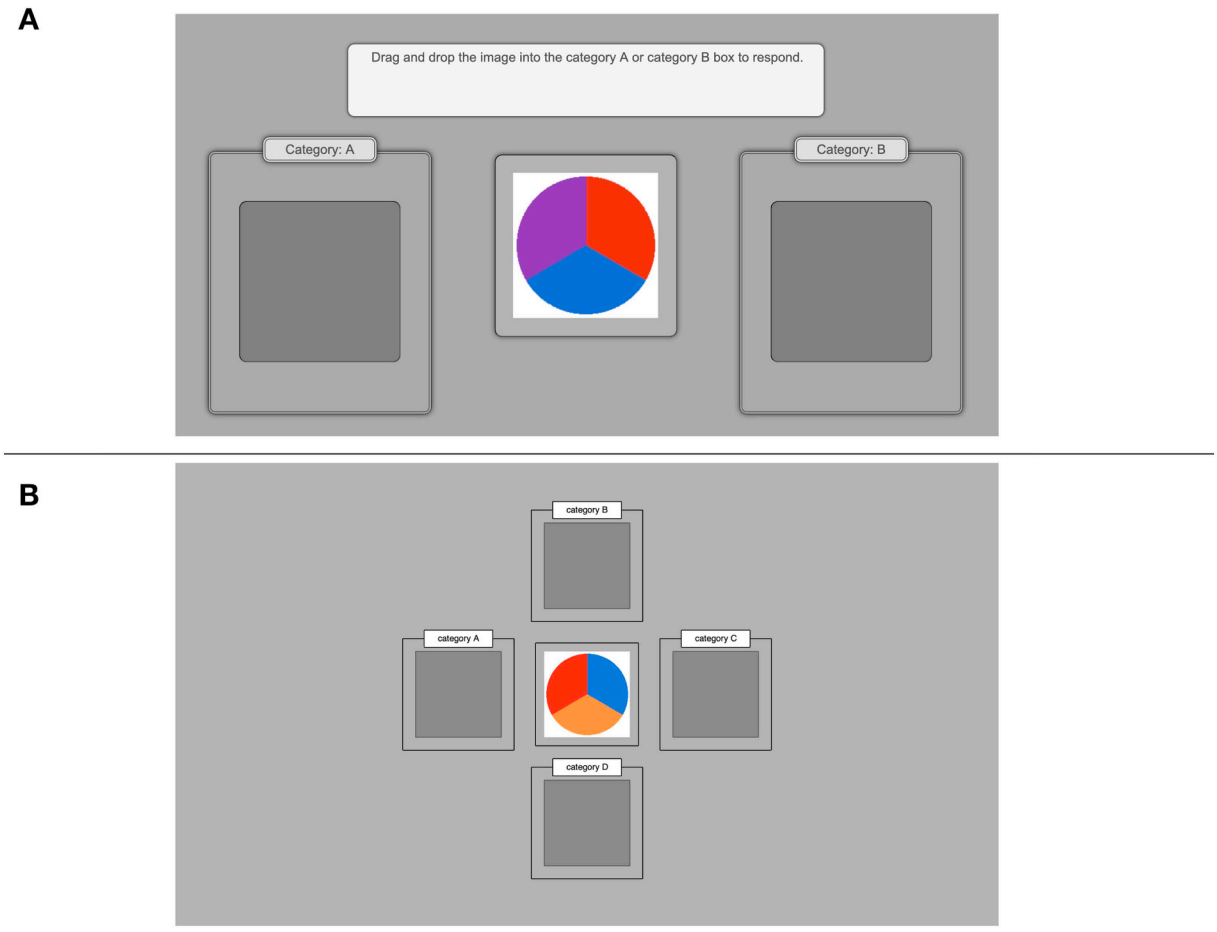

Fig. 3. Example of a trial in (A) Experiments 1A-B and (B) Experiments 2A-B.

predicting participants' trial-by-trial accuracy on training trials from condition, including a by-subject random intercept.[3] We used the lme4 package version 1.1-21 in R (version 3.6.1) to fit all models (D. Bates & Maechler, 2009; R Development Core Team, 2019). Participants in the High Nameability condition (M = 84.0%, 95% CI = [78.6%, 89.4%]) were more accurate than participants in the Low Nameability Condition (M = 67.7%, 95% CI = [59.9%, 75.4%]), b = 1.02, 95% Wald CI = [0.47, 1.56], z = 3.65, p < .001 (see Fig. 4A).

To examine whether participants learned at different rates, we fit a logistic mixed-effects model predicting participants' trial-by-trial accuracy from condition (centered; High Nameability = 0.5, Low Nameability = −0.5), trial number (mean-centered), and their interaction, including a by-subject random intercept and a by-subject random slope for trial number. Accuracy increased across trial number, b = 0.10, 95% CI = [0.07, 0.14], z = 5.67, p < .001, and improved faster in the High Nameability condition than in the Low Nameability condition, b = 0.07, 95% CI = [0.01, 0.14], z = 2.27, p = .023.

Reaction times (RTs) were similar between the two conditions (High: M = 1847 ms, 95% CI = [1676 ms, 2018 ms]; Low: M = 1836 ms, 95% CI = [1650 ms, 2021 ms]) and there was no evidence for a tradeoff between reaction time and accuracy. The reported RT means exclude excessively long reaction times (> 5000 ms.; ~4.4% of trials), but we also did not find any evidence of a speed-accuracy tradeoff for other RT cutoff values.

[3] We did not include random effects for items, since the number of unique items per condition (six) is small and the focus of the research question was not on generalizing across a population of items. However, when by-item random effects are included that properly account for item type (more frequently encountered prototype vs. less frequently encountered non-prototype items), the reported condition effect holds in each experiment.

3. Experiment 1B: learning categories of more and less nameable color features, a replication

In Experiment 1A, categories composed of more nameable colors were easier to learn. In Experiment 1B, we sought to ensure that the results from Experiment 1A were not due to any idiosyncratic properties of the colors in the category-determining position.

3.1. Method

3.1.1. Participants

We recruited 50 new participants (19 female; 49 native speakers of English) through Amazon Mechanical Turk (mean age: 36.0 years; range: 19–69 years). To our knowledge, there is no single standardly used effect size measure for mixed-effects models, so we used the Cohen's d estimated from a two-samples t-test (d = 1.01) as a heuristic for sample size decisions. Given the large effect size in Experiment 1A (93% post-hoc power estimate for n = 50), we reasoned that our sample size would yield sufficient power even if the first experiment overestimated the true effect size. Participants were randomly assigned to the High Nameability Condition (n = 25) or to the Low Nameability Condition (n = 25) and were paid $0.60 for completing the 4-minute task (average completion time: 4.1 min, SD = 1.2).

3.1.2. Stimuli

The stimuli were identical in structure to Experiment 1A and involved the same colors. However, we changed the pairing of the colors

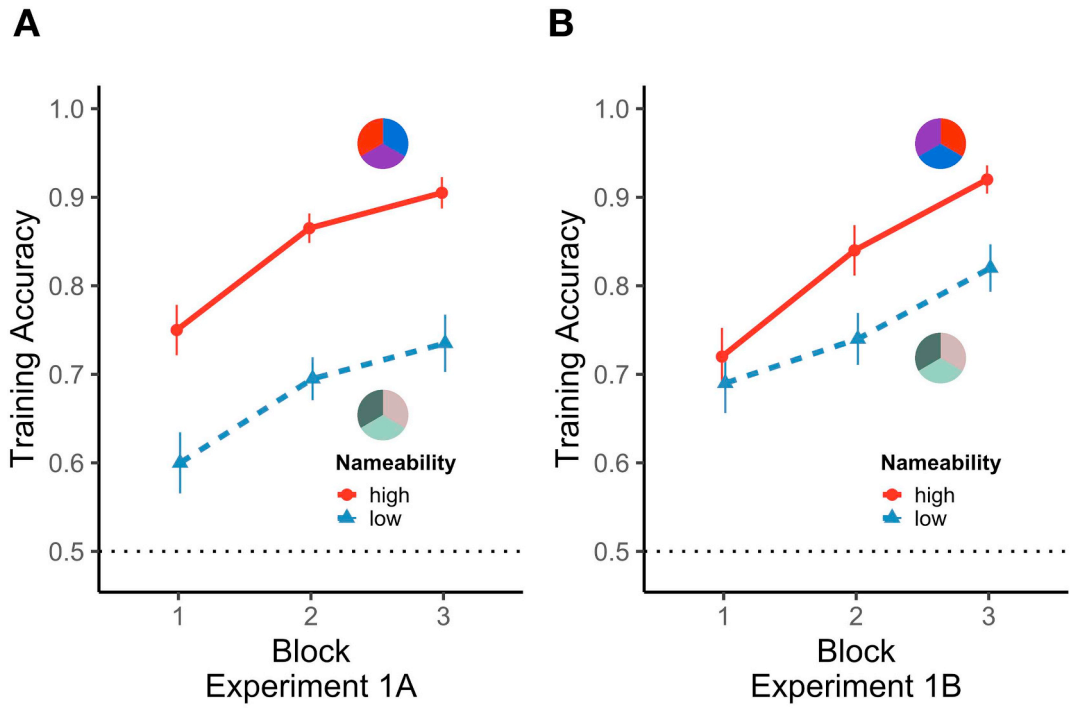

Fig. 4. Accuracy across blocks for (A) Experiment 1A and (B) Experiment 1B. Error bars represent ± 1 SE of the within-subject corrected mean (Morey, 2008).

in the critical position (see Fig. 2B). Although the pairwise ΔE2000 distances of the colors for the prototype stimuli in the two nameability conditions were equated in Experiment 1A, the distance between the two critical colors was slightly higher in the High Nameability condition (ΔE = 56.4) than in the Low Nameability condition (ΔE = 32.1). To rule out the possibility that the effect in Experiment 1A was driven by differences in the discriminability between the two critical colors, in Experiment 1B, we used critical colors in the high nameability condition that had smaller ΔE distance compared to the low nameability condition. The critical colors in the more-nameable categories were "brown" RGB = (120, 80, 40) and "red" RGB = (220, 20, 0) (distance in CIE-LAB space: ΔE = 23.7). The critical colors in the low nameability were the same as in Experiment 1A (RGB = (170,160,40) and RGB = (200,170,170)) (ΔE = 32.1).

3.1.3. Design & procedure

The procedure was identical to that in Experiment 1A.

3.2. Results

3.2.1. Main analyses

We fit the same model as in Experiment 1A to test for differences between conditions. Accuracy was higher in the High Nameability condition (M = 82.7%, 95% CI = [78.2%, 87.2%]) than in the Low Nameability condition (M = 75.0%, 95% CI = [69.3%, 80.7%]), b = 0.48, 95% CI = [0.06, 0.91], z = 2.22, p = .026 (see Fig. 4B). There was no interaction between nameability condition and Experiment (1A vs. 1B), p = .15. To examine whether learning rates differed across experiments, we fit the same logistic mixed-effects model as in Experiment 1, with the exception that the by-subject random slope for trial number was pruned from the model to avoid a singular fit in the random effects covariance matrix (effects are qualitatively similar with or without the by-subject random slope for trial number). As in Experiment 1A, participants' accuracy increased faster in the High Nameability condition than in the Low Nameability condition, b = 0.06, 95% CI = [0.02, 0.11], z = 2.81, p = .005. Reaction times were similar between the two conditions (High: M = 1943 ms, 95% CI = [1771 ms, 2115 ms]; Low: M = 2048 ms, 95% CI = [1794 ms, 2302 ms]) and there was no evidence of a speed-accuracy tradeoff.

3.2.2. Low- vs. high-performing participants in Experiments 1A and 1B

While participants generally succeeded at learning the category structure in Experiments 1A and 1B, we investigated differences in the number of high and low performers in each condition. To do so, we split participants into a "low learner" group and a "high learner" group based on whether a participant achieved at least 75% performance on the final block of the experiment (i.e. was correct on 6 of 8 trials). We find similar results with higher thresholds for low vs. high learners of 7 or 8 correct trials on the final block. We collapsed across Experiments 1A and 1B in our reported results to have more power, given that the set of low learners is relatively small; however, similar trends were found in both experiments considered alone.

Across Experiments 1A and 1B, there were 18 participants (out of 100 total participants) who had < 75% accuracy on the final block ("low learners"). The overwhelming majority of low learners belonged to the Low Nameability condition (16 of 18; p = .001, exact binomial test), suggesting that the category learning task was substantially more difficult in the Low Nameability condition than in the High Nameability condition, despite identical category structure. We further tested whether the effect of nameability held across Experiments 1A and 1B when considering only the set of high learners. Fitting the same main model as in Experiments 1A and 1B, we found a significant effect of condition (b = 0.38, Wald 95% CI = [0.04, 0.71], z = 2.20, p = .028) within the set of high learners alone. Note that this is a far more conservative test of the effect of nameability, given the disproportionate number of participants who did not learn the category structure in the Low Nameability condition.

3.3. Discussion of Experiments 1A and 1B

Experiments 1A-B showed that a simple rule-based category composed of color features is easier to learn when the colors are easier to name. Participants successfully learned the category structure in both conditions, but were more likely to learn the category structure by the final learning block when the category was composed of more nameable colors. Even when participants successfully learned the category,

they were faster to do so when the color features were more nameable. In the Supplementary materials, we also report analyses of participants' verbal response data suggesting that nameability may alter how participants represent the category structure (see Supplementary materials, S2).

Since we are relying on existing differences in nameability, there is always a possibility that the differences in nameability are confounded with another factor that is causing the large difference in accuracy. In Experiment 1A, we equated perceptual discriminability between the features of the prototype exemplars and between the two critical features by matching these features in CIE color space across the high and low nameability conditions. Despite these controls, there remains a possibility that discriminability differences between other pairs of colors could explain the observed differences in accuracy. For example, some of the high nameability color pairs that occurred in the less frequent/non-prototypical category exemplars had particularly high ΔE2000 values (e.g., purple vs. yellow, see Fig. S1 in the Supplementary materials). These differences may have led high nameability category items considered as a set to be more discriminable, which might in turn have made it slightly easier for participants to find the underlying category structure. In addition, the formula for computing CIE-LAB distances may not adequately capture perceptual discriminability as measured through behavior. Aside from discriminability, harder-to-name colors are on the whole less saturated than easy-to-name colors. The strong association between saturation and nameability makes it difficult to derive equi-spaced colors that vary in nameability but do not vary in saturation. Finally, it is possible that participants may have used different strategies for the high and low nameability condition in ways that are difficult to account for in a between-subjects design.

In Experiments 2A and 2B, we sought to address these concerns by (a) constructing a new set of high and low nameable colors that were more equally matched with respect to perceptual discriminability and that were more similar with respect to other color features such as saturation; (b) collecting behavioral norming data on the pairwise perceptual discriminability of the color features, in addition to considering distances in CIE-LAB space; and (c) investigating whether perceptual discriminability differences or other color features such as saturation predicted categorization behavior. To assuage the concern that participants may have used different strategies in the high and low nameability conditions, we manipulated nameability within-subjects. Each participant saw stimuli composed of more nameable colors and stimuli composed of low nameable colors and was tasked with learning to categorize each stimulus into one of four categories.

4. Experiments 2A & 2B: within-subjects color nameability manipulation

In Experiment 2A, we tested the same color set as in Experiment 1B to replicate the results from Experiment 1B in a within-subjects paradigm. In Experiment 2B, we tested a novel stimulus set that was more tightly controlled for perceptual discriminability across all color features. The main analyses for Experiment 2B were pre-registered on OSF (https://osf.io/4euck). The analyses investigating low- and high-performing subjects and testing the effect of color feature characteristics across Experiments 2A and 2B were exploratory.

4.1. Method

4.1.1. Participants

4.1.1.1. Experiment 2A. We recruited 39 University of Wisconsin-Madison undergraduates (23 female; 26 native speakers of English; mean age: 18.5 years; range: 18–20 years) to participate for course credit. One participant was excluded due to a technical malfunction of the response device.

4.1.1.2. Experiment 2B. We recruited an additional 39 University of Wisconsin-Madison undergraduates to participate for course credit (22 female; 31 native speakers of English; mean age: 18.4 years; range: 18–20 years). One participant was excluded due to experimenter error (erroneously scheduling a participant who had participated in a previous color categorization experiment).

4.1.2. Stimuli

4.1.2.1. Color feature selection and norming

4.1.2.1.1. Experiment 2A. To obtain a behavioral measure of the discriminability of the colors used in Experiments 1A, 1B, and 2A, we conducted a norming study in which participants were asked to make speeded same-different judgments about pairs of colors—an extremely sensitive method for measuring represented visual similarity (Lupyan, 2008). 18 new University of Wisconsin-Madison undergraduates participated in the norming task for course credit. In the task, two round color swatches appeared at one of four locations on the screen. Participants were asked to judge as quickly and accurately as possible whether the two colors were the same or different. The color stimuli were the 6 highly nameable and the 6 low nameability colors from Experiments 1A and 1B. High nameability colors were only ever compared to high nameability colors, and low nameability colors only compared to other low nameability colors. The main metric of discriminability was computed over different trials. To increase the number of comparisons between different colors, participants therefore saw twice as many different trials compared to same trials. Each participant saw every possible combination of one color to a different color within a given nameability group a total of 12 times, resulting in a total of 12 * 6 * 5/2 = 180 high nameability different trials and 180 low nameability different trials. In total, participants viewed 540 trials (360 different, 180 same). Color positions and trial order were randomized across participants.

We removed one participant for near chance-level responding (M = 55%) and all trials with reaction times below 200 ms or > 2000 ms (1.7% of trials). Accuracy was extremely high for the remaining participants (M = 97.9%, SD = 1.8%) and similar for high (M = 98.2%) and low nameability colors (M = 97.6%). We computed two discriminability metrics based on the reaction times for correct trials from the resulting data: First, we computed the average reaction time for each individual color across subjects for different trials, which we treated as a general metric of each color's overall discriminability relative to the other colors. To compute this metric, we first calculated the mean reaction time on different trials involving each individual color within each subject, and then averaged across subjects to obtain the discriminability metric (see Table 2 for average reaction times on different trials for each color). Lower reaction times indicate that participants were faster to discriminate a given color from other colors within the set. Although not the main focus of our norming approach, we also computed average reaction times analogously for same trials (see Fig. 5). Second, we computed the average reaction time on different trials for each combination of two colors, first summarizing within- and then across-subjects for each unique combination of two colors. The discussion below focuses on the average reaction times for individual colors. See the Supplementary materials (S1) for an overview of average reaction times obtained for pairwise color combinations, which yield similar conclusions to the average reaction times for individual colors across different trials.

The average reaction times on different trials were similar for high nameability (M = 574 ms, 95% CI = [563 ms, 584 ms]) and low nameability colors (M = 576 ms, 95% CI = [564 ms, 588 ms]), $t(10) = -0.40$, $p = .70$, indicating that participants were able to discriminate high and low nameability colors with similar speed during different trials. However, participants were faster to judge high nameability colors as identical on same trials (M = 598 ms, 95% CI = [589 ms, 608 ms]) than low nameability colors as identical (M = 628 ms, 95% CI = [610 ms, 646 ms]), $t(10) = -3.73$, $p = .004$. This difference seemed to be driven in particular by responses to the

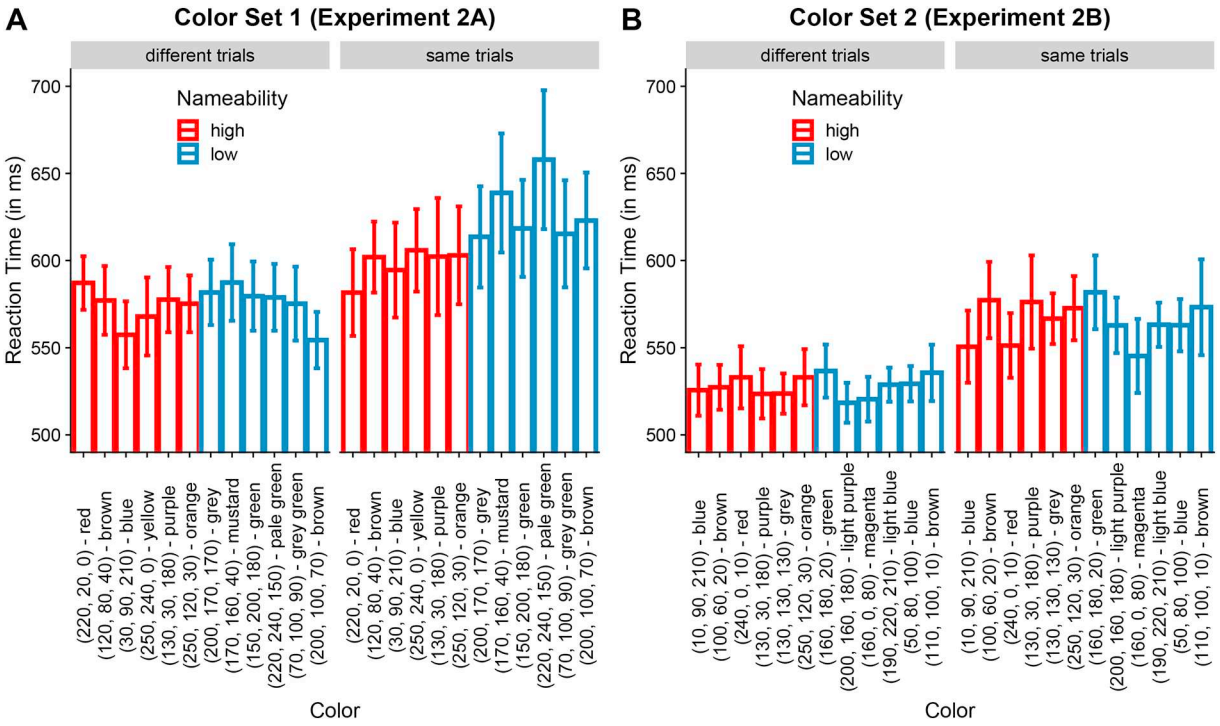

Fig. 5. Average reaction times for each color for both same and different trials for (A) Color Set 1 and (B) Color Set 2. Error bars represent 95% CIs of the within-subject corrected mean (Morey, 2008).

low nameability colors "mustard" RGB = (170,160,40) (M = 639 ms) and "pale green" RGB = (220,240,150) (M = 658 ms). Participants were faster to respond on trials where the two colors were different, presumably because there were twice as many different as same responses, creating a bias to respond "different".

4.1.2.1.2. Experiment 2B. While the color set used in Experiment 2A was selected to control for between-color discriminability for prototype stimuli, there were still differences between colors (measured in terms of ΔE2000) when considering all possible pairwise comparisons between colors in the set (see S1 for an overview of discriminability metrics for all pairwise color comparisons). In Experiment 2B, we selected a new set of high and low nameability colors (Color Set 2) in which we ensured that the ΔE2000 value for every pair of high nameability colors was either virtually identical to (i.e., within 1–2 points on the ΔE2000 metric) or lower (i.e., more difficult to discriminate) than the comparison for the low nameability color pairs that occurred in the analogous role in the category structure (Table 3). For example, the high nameability colors "blue" RGB = (10, 90, 210) and "orange" RGB = (250, 120, 30) selected for Color Set 2 had a ΔE2000 value of 57. The low nameability colors serving an analogous role in the category structure (see Fig. 6B) were "green" RGB = (160, 180, 20) and "magenta" RGB = (160, 0, 80), which had a larger ΔE2000 value of 75 (and were thus more discriminable on the ΔE2000 metric). Note that this method of selecting colors also ensured that the two critical colors (100% predictive of category membership) were selected so that the high nameability colors were

Table 3
Overview over the color feature set in Experiment 2B.

RGB	Color	Modal name	Color nameability	Modal agreement	Simpson diversity	Discriminability RT	Role Exp 2B
(10, 90, 210)		Blue	High	80.7%	0.677	526 ms	100% predictive
(100, 60, 20)		Brown	High	84.4%	0.715	527 ms	100% predictive
(240, 0, 10)		Red	High	85.2%	0.735	533 ms	75% predictive
(130, 30, 180)		Purple	High	82.1%	0.672	524 ms	75% predictive
(130, 130, 130)		Grey	High	78.4%	0.640	524 ms	75% predictive
(250, 120, 30)		Orange	High	85.1%	0.733	533 ms	75% predictive
(160,180,20)		Green	Low	13.3%	0.157	537 ms	100% predictive
(200, 160, 180)		Light purple	Low	10.8%	0.079	518 ms	100% predictive
(50, 80, 100)		Blue	Low	14.7%	0.180	529 ms	75% predictive
(110, 100, 10)		Brown	Low	11.3%	0.111	536 ms	75% predictive
(160, 0, 80)		Magenta	Low	12.7%	0.066	520 ms	75% predictive
(190, 220, 210)		Light blue	Low	14.3%	0.113	529 ms	75% predictive

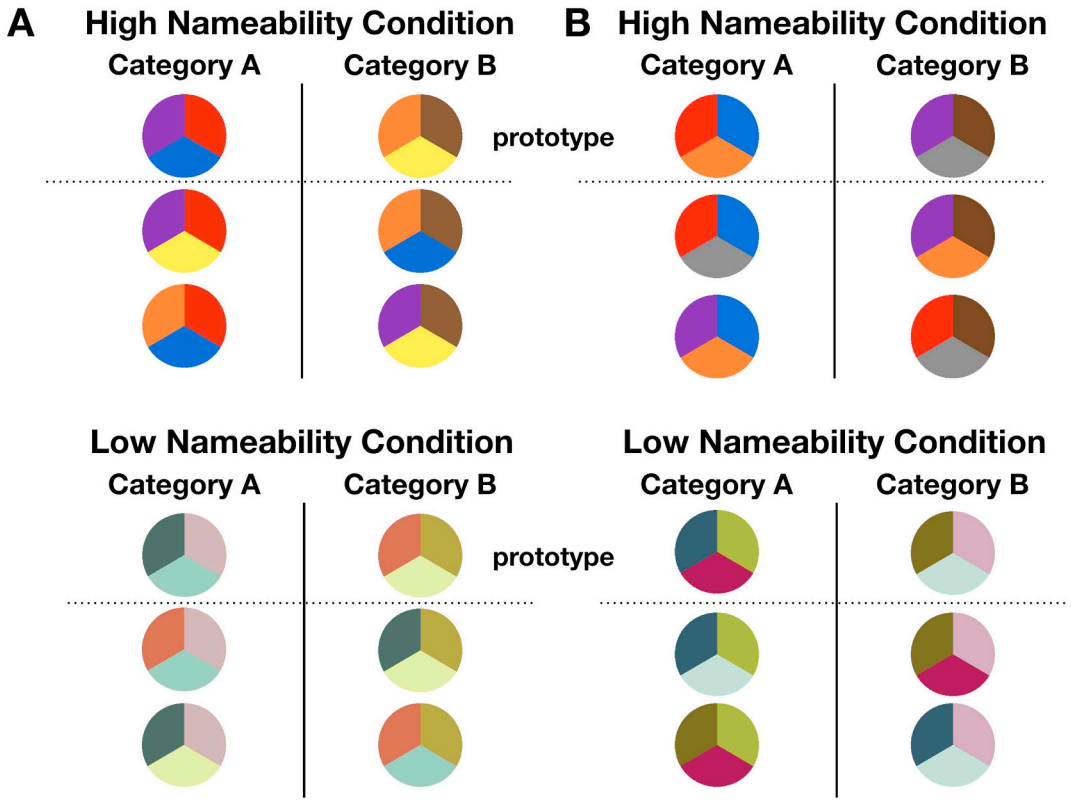

Fig. 6. Stimuli in Experiments 2A-B. Each exemplar (defined by its three color features) could be instantiated with the three colors at any position (e.g., red could occur in the top right, top left or bottom location). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

similar to (and if anything, less discriminable than) the critical colors in the low nameability condition (high nameability colors: ΔE2000 (rgb (10, 90, 210), rgb(100, 60, 20)) = 47; low nameability colors: ΔE2000 (rgb(160, 180, 20), rgb(200, 160, 180)) = 48).

A salient difference between the high and low nameability colors in Color Set 1 was saturation (see Supplementary materials S1, Fig. S3). High nameability colors (M = 82%) were on average more saturated than low nameability colors (M = 44%) in Color Set 1, $t(10) = 3.05$, $p = .01$. While the colors for Experiment 2B were not selected to match on saturation, high nameability (M = 71%) and low nameability colors (M = 59%) were more similar in average saturation in Color Set 2, $t(10) = 0.59$, $p = .57$, though high nameability colors were numerically more saturated on average.

As in Experiment 2A, we conducted a behavioral norming study to measure discriminability of the new color set. The procedure was identical to the behavioral norming study in Experiment 2A, using the new set of high and low nameability colors (Color Set 2). We recruited 20 additional University of Wisconsin-Madison undergraduates who had not taken part in earlier versions of the study. One additional participant was removed for reporting color blindness. Accuracy was high across participants ($M = 97.8\%$, $SD = 1.3\%$) and similar for high (M = 97.9%) and low nameability colors (M = 97.8%).

The average reaction times on different trials were similar for high nameability (M = 528 ms, 95% CI = [523 ms, 532 ms]) and low nameability colors (M = 528 ms, 95% CI = [520 ms, 536 ms]), $t(10) = -0.13$, $p = .90$, indicating that participants were able to discriminate high and low nameability colors equally well. Moreover, participants also judged high nameability colors as identical on same trials (M = 566 ms, 95% CI = [553 ms, 579 ms]) with similar speed compared to low nameability colors (M = 565 ms, 95% CI = [552 ms, 578 ms]), $t(10) = 0.13$, $p = .90$. The high and low nameability colors were therefore closely matched on discriminability, measured both in terms of ΔE2000 values and in terms of the behavioral norming data (see S1 for further details).

Note that participants also responded faster on average in discriminating colors from Color Set 2 compared to Color Set 1 (see Fig. 5). While the source of this difference is unclear, one possible explanation is that the colors in Color Set 2 were easier to discriminate from one another in general, compared to the colors in Color Set 1. Since our main goal was to equate the discriminability of colors within a given set, this between-set difference, while intriguing, is not central to subsequent analyses.

4.1.2.2. Category structure. In Experiment 2A, the stimuli were constructed using the identical color features (Color Set 1) and the same category structure used in Experiment 1B, to allow us to assess whether the original nameability effect would hold even when testing the low and the high nameability stimuli together. The stimuli in Experiment 2B were constructed in an identical manner to the stimuli in Experiment 2A, only using the new set of color features (Color Set 2; see Fig. 6). To ensure that the task was not too easy, we added more variability to the category items by presenting the items randomly in all possible location combinations.

4.1.3. Design & procedure

The procedure in Experiments 2A and 2B was similar to Experiments 1A-1B, with a few central differences. First, nameability was manipulated within participants. Participants were asked to sort both high nameability and low nameability stimuli into one of four different category boxes (top, bottom, left, right, see Fig. 3B). The target category boxes for both the high nameability stimuli and the low nameability stimuli, respectively, were always sorted along the same axis (i.e., if the high nameability stimuli of one category belonged in the top box, then the high nameability stimuli from the other category

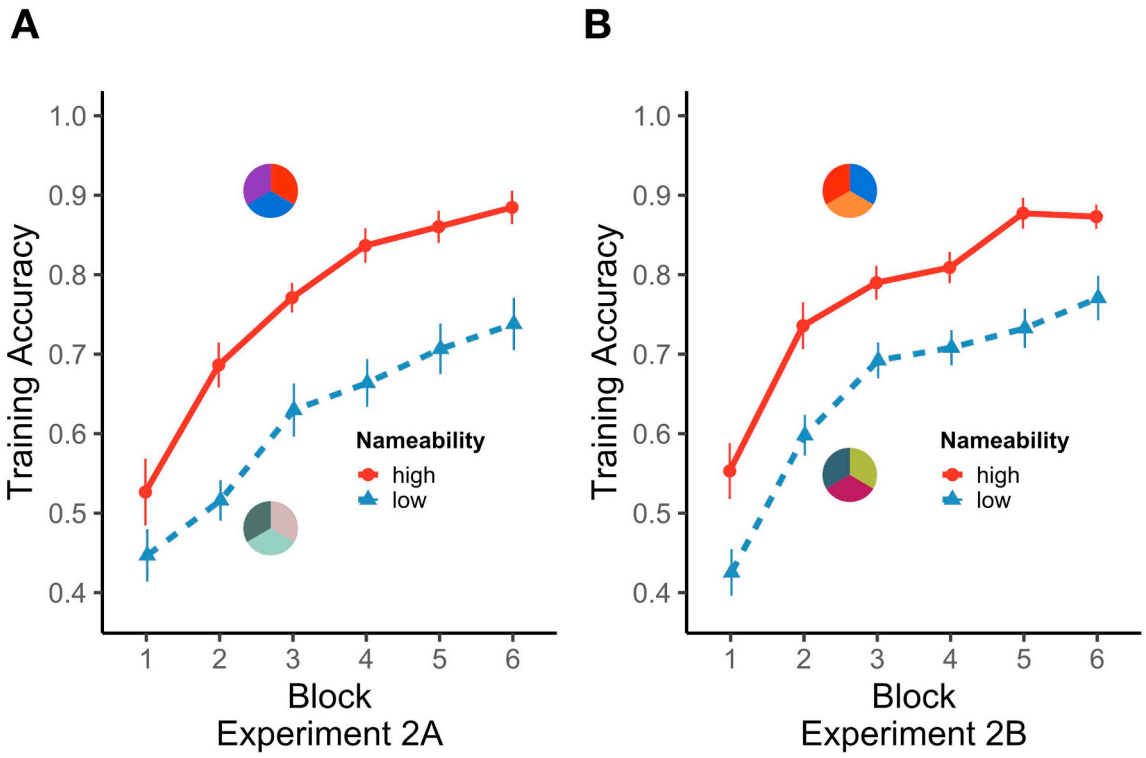

Fig. 7. Accuracy across blocks for (A) Experiment 2A and (B) Experiment 2B. Error bars represent ± 1 SEs of within-subject adjusted means (Morey, 2008).

belonged in the bottom box). The pairing of the high and low nameability stimuli with box locations was counterbalanced across participants. Each block consisted of 16 trials, half belonging to the high nameability condition and half to the low nameability condition. Participants completed 6 blocks for a total of 96 trials. Feedback was identical to Experiments 1A and 1B: participants received immediate feedback on whether their choice was correct or incorrect after each trial and trials were repeated after an incorrect response. The same procedure was used in both Experiments 2A and 2B – the only difference between experiments was in the stimulus set participants were tested on (see Fig. 6).

4.2. Results

4.2.1. Experiment 2A

We tested whether accuracy differed between conditions using the same models as in Experiment 1A, with the exception that we included a by-subject random slope for condition, given that condition was now a within-subject factor. Trials with very short (< 200 ms) and very long (> 5000 ms) response times were removed (2.2% of trials). All results below remain similar if these trials are retained for analysis. Accuracy was higher in the High Nameability condition (M = 76.3%, 95% CI = [71.3%, 81.3%]) than in the Low Nameability condition (M = 61.9%, 95% CI = [57.0%, 66.8%]), b = 0.80, 95% CI = [0.54, 1.06], z = 5.99, p < .001 (see Fig. 7A). Participants' accuracy increased across trials (b = 0.028, 95% CI = [0.021, 0.034], z = 8.43, p < .001), and increased more rapidly in the High Nameability condition than in the Low Nameability condition, b = 0.014, 95% CI = [0.008, 0.020], z = 4.27, p < .001. In addition to being more accurate, stimuli containing more nameable colors were also categorized more quickly (High: M = 1137 ms, 95% CI = [1047 ms, 1226 ms]; Low: M = 1342 ms, 95% CI = [1235 ms, 1449 ms]). However, there was no evidence of a speed-accuracy tradeoff.

4.2.2. Experiment 2B

We fit the same models as in Experiment 2A to test for differences between conditions. Trials with very short (< 200 ms) and very long (> 5000 ms) response times were removed (2.2% of trials). All results below remain similar if these trials are retained for analysis. Accuracy was higher in the High Nameability condition (M = 77.6%, 95% CI = [72.4%, 82.8%]) than in the Low Nameability condition (M = 65.5%, 95% CI = [60.7%, 70.3%]), b = 0.72, 95% CI = [0.48, 0.96], z = 5.88, p < .001 (see Fig. 7B). Accuracy improved across trials overall, b = 0.026, 95% CI = [0.020, 0.032], z = 8.71, p < .001, and increased more rapidly in the High Nameability condition than in the Low Nameability condition, b = 0.007, 95% CI = [0.0005, 0.013], z = 2.11, p = .035. More nameable stimuli were categorized marginally more quickly than less nameable stimuli (High Nameability: M = 1234 ms, 95% CI = [1134 ms, 1333 ms]; Low Nameability: M = 1357 ms, 95% CI = [1264 ms, 1449 ms]).

4.2.3. Low- vs. high-performing participants in Experiments 2A and 2B

As in Experiments 1A-1B, we investigated differences in the number of high and low performers in each condition, collapsing across Experiment 2A and 2B. We considered the number of participants who were < 75% accurate on the final block (i.e., responded correctly on fewer than 6 of the 8 trials) for low nameability trials, high nameability trials or both ("low performers"). Overall, 8 participants (10.3%) were < 75% accurate on the final block for both low and high nameability trials, 17 participants were low performers for low nameability trials only, and 3 participants were low performers on high nameability trials only. An exact multinomial test showed that the distribution of participants across these three categories was significantly skewed (p = .004), with more low performers on low nameability trials. We also considered whether the nameability effect in Experiments 2A and 2B held when considering high performers alone. There was a similar effect of condition after removing low performers (i.e., participants who were < 75% accurate on either low or high nameability trials – or both – in the final block), b = 0.70, 95% Wald CI = [0.50, 0.90], z = 6.78, p < .001.

4.2.4. Overall analyses including color feature characteristics

Next, we conducted an exploratory analysis asking what features of the category exemplars best predicted categorization performance across both Experiment 2A and Experiment 2B: (a) perceptual discriminability of the color features based on the ΔE2000 values, (b) perceptual discriminability of the color features of each stimulus based on the behavioral norming data (i.e. reaction times), (c) perceptual discriminability of the two 100% predictive (category-determining) color features within a given condition (based on the behavioral norming data) (d) the saturation of the color features (a salient difference between the colors in set 1 and – to a lesser extent – in set 2) or (e) the nameability of the color features (measured continuously based on Simpson's diversity metric). Note that the ΔE2000 values were not included in the model, since we specifically selected the 100% predictive colors to be slightly less discriminable in the high nameability condition than in the low nameability condition based on this metric.

We fit a logistic mixed-effects model predicting participants' accuracy on each trial across both Experiment 2A and Experiment 2B from the average pairwise ΔE2000 discriminability of the three colors of a given stimulus (centered within participants), the average discriminability of the stimulus colors based on pairwise reaction times from the behavioral norming studies (z-scored within each color set), the average discriminability of the two 100% color features within the exemplar's condition (z-scored within each color set), the average saturation of the three colors of the tested category exemplar (centered within participants), and the average Simpson's diversity index of the three colors (centered within participants), while controlling for stimulus type (prototype vs. non-prototype; centered). We included by-participant and by-stimulus random intercepts and by-participant random slopes for average Simpson's diversity index, average within-stimulus color reaction time discriminability, and average saturation after pruning the maximal random effects structure due to non-convergence. Reaction times were z-scored within each experiment given our interest in the relative discriminability of color pairs within each set. However, alternative ways of transforming the data (e.g., centering within participants) yield qualitatively equivalent results.

Average saturation ($b = -0.31$, 95% CI = [−0.96, 0.35], $z = -0.92$, $p = .36$) and the average reaction time discriminability of the 100% predictive color features ($b = -0.007$, 95% CI = [−0.27, 0.25], $z = -0.05$, $p = .96$) were not related to participants' accuracy. Both higher average ΔE2000 discriminability ($b = 0.012$, 95% CI = [0.002, 0.022], $z = 2.37$, $p = .02$) and lower average pairwise reaction time differences (i.e., higher average pairwise discriminability based on the behavioral norming data; $b = -0.35$, 95% CI = [−0.56, −0.15], $z = -3.42$, $p < .001$) predicted higher accuracy overall. Crucially, higher average nameability (Simpson's diversity index) predicted higher accuracy, $b = 1.45$, 95% CI = [0.72, 2.19], $z = 3.87$, $p < .001$, after controlling for pairwise discriminability within each stimulus (both based on ΔE2000 and the behavioral norming data), the discriminability of the two 100% predictive colors in a given condition, saturation, and stimulus type.

4.3. Discussion

Experiments 2A-2B replicated and extended the results from Experiments 1A-1B while addressing several potential limitations of the initial experiments. Replicating the effects from Experiments 1A-1B in a within-subjects design (both with the same color set used in Experiments 1A-1B and a novel color set) allowed us to rule out that the main source of the nameability effect came from some general difference in how learners engaged with the task when seeing high versus low nameability stimuli. In the within-subjects design, participants completed the same category learning task simultaneously for both the high and the low nameability items. In fact, given that learners could leverage their learning with the high nameability items to improve their accuracy on the low nameability items (e.g., recognizing the general rule that a single color feature predicted category membership), one might expect that this design should reduce the size of nameability effects. Nevertheless, we found robust differences in category learning between the high and the low nameability items.

Most importantly, Experiments 2A and 2B allowed us to investigate whether differences in color discriminability not accounted for in Experiments 1A and 1B could explain the observed condition differences. First, we collected behavioral norming data as a second source of information on color discriminability. Second, we tested a novel color set in Experiment 2B that was carefully matched in discriminability (both in terms of ΔE2000 values and in terms of the behavioral norming data) across all within- and across-exemplar color feature comparisons, ensuring that all pairwise low nameable color comparisons were either equally or more discriminable than their high nameable color counterparts. We obtained similar results using this new color set in Experiment 2B, suggesting that the effect of nameability is robust across various perceptual discriminability metrics. Moreover, the variation in color features across Experiments 2A and 2B allowed us to test which factors best explained category learning performance. Across experiments, nameability was a robust predictor of category learning accuracy, even after controlling for perceptual characteristics such as the average discriminability or saturation of color features. It is notable that perceptual discriminability as measured through the behavioral reaction time experiments and ΔE2000 values was also predictive of category learning accuracy. This suggests that feature discriminability impacts participants' ability to learn novel categories, perhaps because higher feature discriminability makes it easier to isolate individual features. However, feature nameability was a robust predictor over and above any effects of feature discriminability.

5. Experiment 3A: learning categories of more and less nameable shape features

In Experiments 3A-B, we extended the nameability effect from Experiments 1A-B and 2A-B to the domain of shapes. By investigating a new domain, we aimed to test the generalizability of the effect of nameability on category learning across different types of stimuli. Moreover, testing category learning in the shape domain allowed us to circumvent some of the difficulties of controlling perceptual metrics in the color domain. In Experiment 3A, we first tested whether items constructed from more nameable shapes lead to better accuracy in a design identical to Experiments 1A-B.

5.1. Method

5.1.1. Participants

We recruited 48 participants (20 female; all native speakers of English) through Amazon Mechanical Turk (mean age: 36.3 years; range: 18–65 years). Data analysis began after 50 participants submitted their work on Amazon Mechanical Turk. However, two participants' data were not submitted to our server (either because the participants did not complete the study for unknown reasons or due to an unknown technical error). Based on the Cohen's d estimated from experiments 1A-1B ($d = 0.83$), $n = 48$ corresponded to 80% power to detect the condition effect. Participants were randomly assigned to the High Nameability Condition ($n = 23$) or to the Low Nameability Condition ($n = 25$) and were paid $0.80 (average completion time: 5.7 min, $SD = 1.6$).

5.1.2. Stimuli

The exemplars were circles (“shape wheels”) analogous to the stimuli used in Experiments 1A-B, except composed of 3 different shapes instead of colors (see Fig. 8A). The structure of the stimuli and categories was identical to Experiments 1A-B: one of the shapes was perfectly predictive of category membership, while the other two were correlated at 75% with category membership.

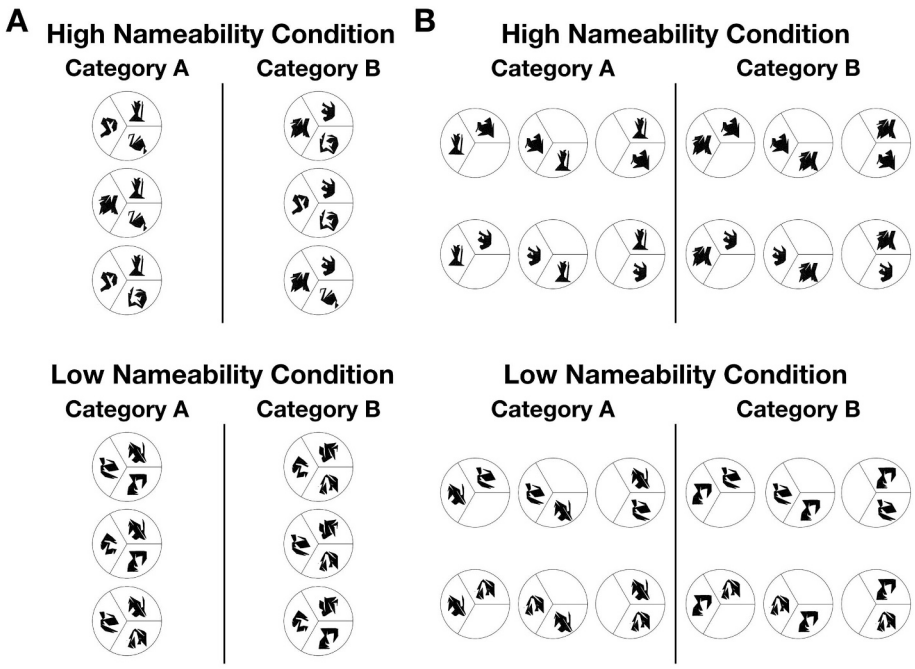

Fig. 8. Stimulus structure in (A) Experiment 2A and (B) Experiment 2B.

The shapes were chosen from a study on the meaningfulness of randomly generated shapes (Vanderplas & Garvin, 1959). In the original study, the shapes were created by randomly generating points and connecting them to form irregular polygons with a fixed number of corners. These stimuli were then normed based on the consistency with which participants attempted to give a one- to two-word description of the shape. A shape's "association value" was defined as the percent of participants who either named the shape or indicated that the shape reminded them of something. Note that this "association value" measure is related to but not identical to our operationalization of nameability for the color stimuli in Experiment 1 (see below for a discussion of participants' naming data). We used these association values in constructing the stimulus set under the assumption that the association values would correlate highly (but not necessarily perfectly) with nameability.

From the 24-point shapes, we chose the six shapes with the highest association value as the items for the High Nameability condition, and the six shapes with the lowest association value as the items for the Low Nameability condition (Fig. 8A). Based on nameability data collected post-hoc in Experiment 3B, the items in the High Nameability condition were more nameable than the items in the Low Nameability condition (High: Average Simpson's Diversity Index = 0.132, Average Modal Agreement = 29.5%; Low: Average Simpson's Diversity Index = 0.016, Average Modal Agreement = 9.8%), though there was some overlap in the distribution of nameability between the two sets (see Experiment 3B for a more extended discussion of the relationship between association values and our nameability measures).

To ensure that the set of shapes in the High Nameability condition and in the Low Nameability condition were matched on overall shape complexity, we estimated the shape skeleton for each image using the method described in Feldman and Singh (2006). In a shape skeleton, a shape contour is represented as a set of curves that follow the shape structure. The complexity of the optimal shape skeleton is captured by the shape skeleton's description length, which corresponds to how difficult it is to encode the best-fitting skeleton (Feldman & Singh, 2006). The average description length for the shape skeletons was similar in the High Nameability condition (Average Skeleton Description Length = 5409, 95% CI = [4094, 6724]) and the Low Nameability condition (Average Skeleton Description Length = 5444, 95% CI = [5082, 5806], $t(10) = -0.07$, $p = .95$).

5.1.3. Design & procedure

The procedure was identical to that in Experiment 1A.

5.2. Results

The analytic approach was identical to Experiment 1A. Participants in the High Nameability condition ($M = 85.9\%$, 95% CI = [80.6%, 91.1%]) outperformed participants in the Low Nameability condition ($M = 68.3\%$, 95% CI = [60.8%, 75.9%]), $b = 1.15$, 95% CI = [0.58, 1.72], $z = 3.95$, $p < .001$ (see Fig. 9A). Accuracy improved across trials overall, $b = 0.19$, 95% CI = [0.12, 0.26], $z = 5.37$, $p < .001$, and increased more rapidly in the High Nameability condition than in the Low Nameability condition, $b = 0.17$, 95% CI = [0.05, 0.29], $z = 2.83$, $p = .005$. There was no evidence for a speed-accuracy tradeoff (Average reaction times in High Nameability Condition: $M = 1820$ ms, 95% CI = [1609 ms, 2031 ms]; Low Nameability Condition: $M = 1955$ ms, 95% CI = [1752 ms, 2158 ms]).

6. Experiment 3B: learning categories of more and less nameable shape features, a replication

In Experiment 3A, categories composed of more nameable shapes were learned more quickly than categories composed of less easily named shapes. However, although the shapes were equated on coarse perceptual measures such as the number of points, the ease of visually discriminating between pairs of shapes in the high and low nameability conditions was not controlled. The shapes in the High Nameability condition might be perceptually easier to distinguish compared to shapes in the Low Nameability condition, which might explain better category-learning performance. In Experiment 3B, we replicated our findings from Experiment 3A while controlling for shape discriminability.

6.1. Method

6.1.1. Participants

We recruited 120 participants (58 female; all native speakers of

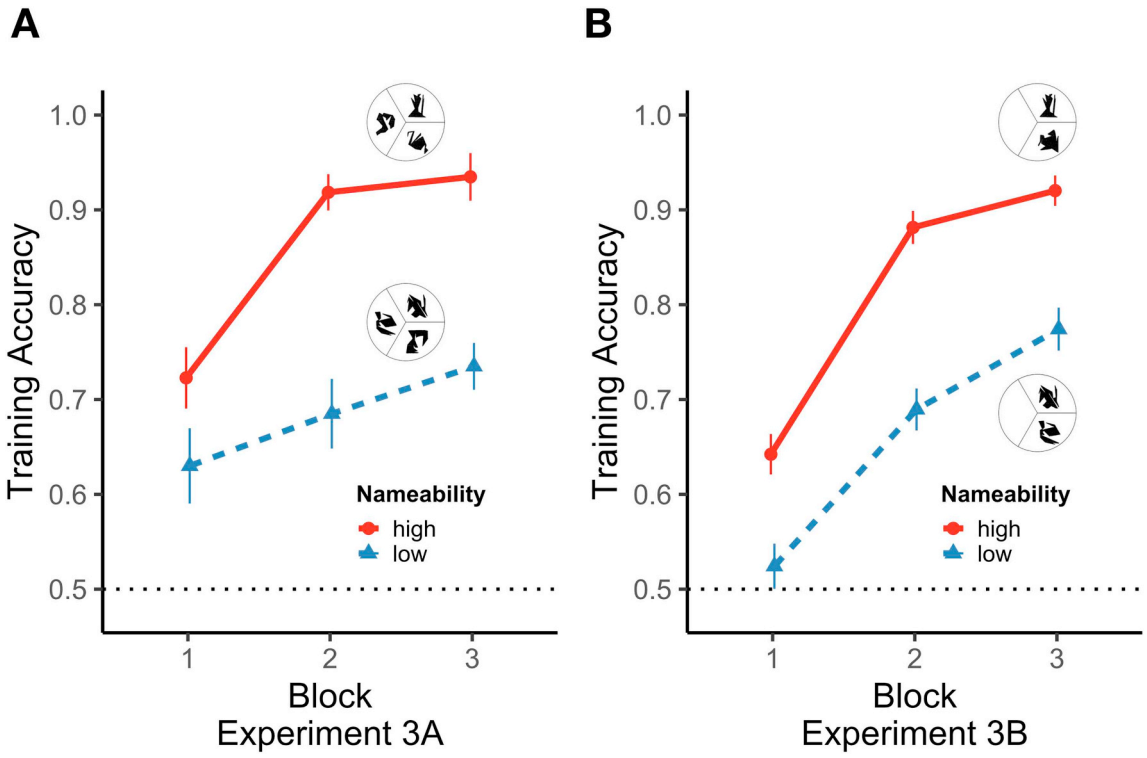

Fig. 9. Accuracy across blocks for (A) Experiment 3A and (B) Experiment 3B. Error bars represent +1/−1 within-subjects SEs (Morey, 2008).

English) through Amazon Mechanical Turk (mean age: 34.6 years; range: 20–63 years). We increased the sample size to have sufficient power even if simplifying the experiment structure (as described below) reduced the effect size. Participants were randomly assigned to the High Nameability Condition (n = 58) or to the Low Nameability Condition (n = 62) and were paid $0.90 for completing the task, which lasted approximately 8 min (average completion time: 8.0 min, SD = 2.4).

6.1.2. Stimuli

6.1.2.1. Stimulus norming and selection. To ensure that shapes used in the experiment were equally discriminable, we conducted a separate in-lab norming task in which participants (n = 22) performed a speeded same/different task as in Experiments 2A and 2B. We collected visual discriminability data for the eight most nameable and the eight least nameable shapes from Vanderplas and Garvin's ratings using the same procedure as in Experiments 2A and 2B. We also collected naming responses from all 8 high and 8 low nameability shapes at the end of the norming task (see Table 4 for the resulting modal agreement and Simpson diversity values).

In general, the high nameability shapes were slightly more discriminable than the low nameability shapes. Participants were slightly more accurate for the 8 high nameability shapes (M = 97.5%, 95% CI = [97.0%, 98.0%]) than for the low nameability shapes (M = 96.0%, 95% CI = [95.5%, 96.5%]). Participants were also substantially faster in correctly discriminating the high nameability shapes (different trials: M = 626 ms, 95% CI = [617 ms, 635 ms]; same trials: M = 670 ms, 95% CI = [651 ms, 689 ms]) than discriminating the low nameability shapes (different trials: M = 699 ms, 95% CI = [682 ms, 715 ms]; same trials: M = 710 ms, 95% CI = [685 ms, 734 ms]), leaving an insufficient number of shapes for creating the same feature structure as in Experiment 2A while still matching the two conditions on shape discriminability. We therefore selected four highly nameable shapes and four shapes with low nameability such that differences in pairwise discriminability (as measured by reaction times in the norming task) were matched as closely as possible between the two shape sets (high nameability shapes: M = 652 ms, 95% CI = [617 ms, 688 ms]; low nameability shapes: M = 665 ms, 95% CI = [633 ms, 697 ms]; t(10) = −0.70, p = .50), The average pairwise discriminability ranged from 622 ms - 703 ms for high nameability shapes and from 617 ms–692 ms for the low nameability shapes. The two shapes that were 100% predictive of category membership were the two shapes with the lowest reaction time values (high: 623 ms; low: 616 ms), i.e., the two shapes that were most distinguishable (see Supplementary materials, Table S3 and S4 for further details). The four items in the High Nameability condition were similar to the four items in the Low Nameability condition in shape skeleton complexity (High Nameability Average Description Length = 5885, 95% CI = [3410, 8360]; Low Nameability Average Description Length = 5423, 95% CI = [4948, 5898]; t(6) = 0.58, p = .58), with the shapes in the High Nameability condition being slightly *more* complex than the shapes in the Low Nameability condition.

In general, the shapes grouped as high nameability items were higher on our measures of nameability than the shapes grouped as low nameability items (High Average Modal Agreement = 26.7%, Low Average Modal Agreement = 10.2%, t(14) = 2.34, p = .03; High Nameability Average Simpson's Diversity Index = 0.117, Low Nameability Average Simpson's Diversity Index = 0.014, t(14) = 2.17, p = .047). There was a medium-sized correlation between the association value of a shape (from Vanderplas and Garvin) and its nameability, measured in terms of Simpson's diversity index (r = 0.54, p = .032) and modal agreement (r = 0.55, p = .028). Thus, association value was a reliable but imperfect indicator of nameability in Experiment 3A, leading to some overlap in the nameability scores of the shapes grouped into the high and low nameability sets. In Experiment 3B, the four shapes selected for the high nameability set (Average Modal Agreement = 30.7%; Average Simpson's Diversity Index = 0.14) were in general more nameable than the four shapes in the low nameability set (Average Modal Agreement = 9.1%; Average Simpson's Diversity Index = 0.01). All four of the high nameability shapes scored higher or equal to the highest nameability scores among the low nameability items on both modal agreement and Simpson's diversity index.

Table 4
Overview of the shape feature set in Experiments 3A & 3B.

Shape	Association value (Vanderplas & Garvin, 1959)	Name-ability	Modal name	Modal agreement	Simpson diversity	Shape skeleton description length	Role Exp 3A	Role Exp 3B
	60%	High	Tree	13.6%	0.033	4655	100% predictive	100% predictive
	54%	High	Bear	59.1%	0.345	4452	100% predictive	50% predictive
	48%	High	Swan	54.5%	0.304	6388	75% predictive	–
	48%	High	Lizard	9.1%	0.008	5039	75% predictive	–
	44%	High	Heart	18.2%	0.046	4432	75% predictive	–
	42%	High	Duck	22.7%	0.054	7488	75% predictive	100% predictive
	42%	High	Frog	27.3%	0.126	6944	–	50% predictive
	42%	High	Letter E	9.1%	0.017	5337	–	–
	22%	Low	Frog	13.6%	0.011	5624	100% predictive	100% predictive
	24%	Low	Bird	9.1%	0.014	5886	100% predictive	–
	28%	Low	Claw	4.5%	0.002	5001	75% predictive	100% predictive
	28%	Low	Dolphin	9.1%	0.009	5421	75% predictive	50% predictive
	28%	Low	Angel	9.1%	0.019	5646	75% predictive	50% predictive
	28%	Low	Finger pointing	13.6%	0.038	5084	75% predictive	–
	30%	Low	Face	9.1%	0.006	6318	–	–
	32%	Low	Bat	13.6%	0.011	5195	–	–

Note. Shapes not used in a given experiment are marked as "–" in the category role columns.

6.1.2.2. Stimulus construction. We created four high nameability two-shape combinations and four low nameability two-shape combinations using the four high nameability shapes and the four low nameability shapes matched on perceptual discriminability (see Fig. 8B). As in Experiment 2A, the categories were defined by the presence of a single critical shape for both the high nameability and the low nameability items (termed the 100% predictive shape). Within a given condition, each two-shape combination was composed of one of two possible 100% predictive shapes and one of two possible 50% predictive shapes, resulting in two unique combinations of two shape features belonging to each category. The use of just two shapes per stimulus substantially simplified the category structure, leading to a worry that the categories would be too easy to learn. We therefore sought to increase difficulty by randomizing the location of the shapes to create more within-category variability. This meant that the category-defining critical shape appeared in all three "slice" locations rather than in a fixed location, as in Experiment 3A. For each two-shape combination, there are six possible location arrangements across the three shape wheel locations. For each participant, a subset of three of these locations were selected at random for each of the four two-shape combinations. Thus, in both the high nameability and the low nameability condition, each category was composed of six unique stimuli, created by arranging two combinations of two shape features into unique location configurations in the shape wheel for each participant (see Fig. 8B for a specific example of a resulting stimulus set).

6.1.3. Design & procedure

The trial procedure was identical to Experiment 3A using the novel

two-shape category items (see Fig. 8B). Each stimulus was seen twice, for a total of 24 trials. To maintain consistency in the presentation of the results, we group these 24 trials into 3 blocks of 8 trials in visualizing the results.

6.2. Results

6.2.1. Category learning

Participants in the High Nameability condition (M = 81.5%, 95% CI = [77.7%, 85.3%]) outperformed participants in the Low Nameability condition (M = 66.3%, 95% CI = [61.7%, 70.8%]), b = 0.89, 95% CI = [0.55, 1.23], z = 5.17, $p < .001$ (see Fig. 9B). Experiment (3A vs. 3B) did not interact with the nameability factor, p = .45. Accuracy improved across trials overall, b = 0.19, 95% CI = [0.16, 0.23], z = 10.68, $p < .001$, and increased more rapidly in the High Nameability condition than in the Low Nameability condition, b = 0.12, 95% CI = [0.05, 0.18], z = 3.69, $p < .001$. Lower accuracy corresponded to generally slower RTs (Average reaction times in High Nameability Condition: M = 1862 ms, 95% CI = [1746 ms, 1978 ms]; Low Nameability Condition: M = 1931 ms, 95% CI = [1804 ms, 2058 ms]).

6.2.2. Low- vs. high-performing participants in Experiments 3A and 3B

As in Experiments 1A and 1B, we investigated differences in the number of high and low performers in each condition across Experiments 3A and 3B by splitting participants into a “low learner” group and a “high learner” group based on whether that participant achieved at least 75% performance on final block of the experiment (i.e. was correct on 6 of last 8 trials; we find similar results with higher thresholds of 7 or 8 correct trials on the final block). Across Experiments 3A and 3B, there were 47 participants (28% of all participants) who had < 75% accuracy on the final block (“low learners”). Significantly more low learners belonged to the Low Nameability condition than the High Nameability condition (36 of 47; $p < .001$, exact binomial test). Moreover, we replicated the effect of nameability when considering only the high-performing group of participants, finding an effect of condition when fitting the same main model as in Experiments 3A and 3B (b = 0.58, Wald 95% CI = [0.29, 0.88], z = 3.90, $p < .001$).

6.2.3. Overall analyses including shape feature characteristics

To assess the relative contributions of perceptual discriminability and item nameability to categorization accuracy, we fit a logistic mixed effects model predicting trial-by-trial accuracy across Experiments 3A and 3B from the average discriminability of the shape features of the tested category exemplar based on pairwise reaction times from the behavioral norming study (z-scored), the average complexity (shape skeleton description length) of the shape features (z-scored), and the average Simpson's diversity index of the shape features (i.e., their nameability). We also included the perceptual discriminability of the two 100% predictive (category-determining) shape features within a given condition (based on the behavioral norming data) as a fixed effect. We included the maximal random effects structure that still allowed the model to converge (Barr, Levy, Scheepers, & Tily, 2013), including by-participant and by-item random intercepts, as well as a by-participant random slope for average Simpson's diversity index and for average perceptual discriminability of the within-stimulus shape features.

The average nameability of shape features predicted higher accuracy across Experiments 3A and 3B while controlling for within-exemplar shape discriminability, within-exemplar description length, and the discriminability of the 100% predictive shape features, b = 4.84, Wald 95% CI = [2.49, 7.19], z = 4.04, $p < .001$. The average within-exemplar discriminability of shape features (b = −0.003, Wald 95% CI = [−0.14, 0.13], z = −0.05, p = .96), the average description length of shape features (b = −0.01, Wald 95% CI = [−0.14, 0.11], z = −0.23, p = .82), and the discriminability of the 100% predictive shape features in a given condition (b = −0.02, Wald 95% CI = [−0.17, 0.12], z = −0.33, p = .74) were unrelated to participants' categorization accuracy. Thus, the nameability of shape features (as measured by Simpson's diversity index) was a robust predictor of participants' accuracy across conditions, controlling for shape feature discriminability and complexity.

6.3. Discussion

In Experiments 3A and 3B, we tested whether the advantage shown by high nameability colors in Experiments 1–2 can be observed in another domain: easier- vs. harder-to-name shapes. In Experiment 3A, we tested the hypothesis that participants would learn categories composed of easier-to-name shapes better than categories composed of harder-to-name shapes by using novel polygons that were matched on complexity but varied in the degree to which they evoked meaningful labels. In Experiment 3B, we further controlled for perceptual discriminability between easier- and harder-to-name shapes, obtaining similar results. Moreover, when combining across the results from both studies, we found that the nameability of the shapes was a highly robust predictor of categorization accuracy even after controlling for differences in perceptual discriminability and an objective measure of perceptual complexity.

One reason why participants in the high nameability condition perform more accurately may be because they are more successful at forming (verbal) hypotheses about individual features of the category exemplars. In the Supplementary materials (S2, Fig. S4), we provide further evidence for this view from participants' self-reported verbal strategies. In general, participants in the high nameability condition were more likely to represent the category structure in terms of individual features (either a single feature, e.g. “the one with the bear goes to the left”, or multiple features, e.g. “the ones with the tree and the swan go to the right”) as compared to participants in the low nameability condition. Being able to easily form hypotheses about the relevant category features may allow learners in the high nameability condition to more quickly test different possible hypotheses about the category structure.

Although nameability was a strong predictor of categorization performance in the current experiments, the high nameability shapes used in Experiments 3A and 3B were notably less nameable than the colors used in Experiments 1A-2B, both in terms of modal agreement and in terms of Simpson's diversity index. One reason for the lower nameability scores (in particular for the diversity index) is that shape naming tends to be less constrained and more open-ended than color naming, where there is typically a smaller set of color terms that participants might use. In a direct comparison of a highly nameable color and a highly nameable shape, we would predict that colors should typically receive higher scores. Indeed, in their large-scale study of cross-cultural codability, Majid et al. (2018) found that colors tend to be more nameable than shapes, though there is substantial variability in nameability with respect to both color and shape across languages, and these results may strongly depend on the choice of materials. For the current experiments, the important finding is that within a given feature domain (in this case, color or shape), the relative nameability of features dramatically impacts how easily simple categories are learned.

7. Experiment 4: Manipulating shape nameability through simple rotation

One drawback to the stimuli used in Experiments 3A-B is that more nameable items also tended to be more easily discriminable, limiting our ability to construct stimulus sets fully controlling for discriminability. Although discriminability did not predict categorization accuracy independently of nameability in Exps 3A – 3B, we sought to additionally ensure that differences in nameability affect categorization

Table 5
Overview over the shape feature set in Experiment 4.

Original image	Rotated image	Modal name original	Modal agreement original	Modal name rotated	Modal agreement rotated	Simpson's diversity original	Simpson's diversity rotated	Role Exp 4
		Duck	36.4%	High heel	9.1%	0.140	0.006	100% predictive shape
		Camel	36.4%	Swan	4.5%	0.120	0.005	100% predictive shape
		Cat	31.8%	Fish	13.6%	0.143	0.027	75% predictive shape
		Bird	31.8%	Bat	9.1%	0.128	0.022	75% predictive shape
		Vulture	27.3%	Boat	9.1%	0.095	0.003	75% predictive shape
		Prayer	27.3%	Dolphin	9.1%	0.055	0.005	75% predictive shape

accuracy using shapes that are better controlled for perceptual complexity and discriminability. In Experiment 4, we generalized our findings to new shapes more closely controlled for perceptual complexity and discriminability while manipulating nameability. Instead of selecting different images for the high and low nameability conditions, we manipulated nameability by rotating nameable shapes to produce a harder-to-name orientation. This allows us to manipulate nameability while equating the shapes on all (orientation-independent) perceptual metrics.

7.1. Method

7.1.1. Participants

We recruited 119 participants (49 female; 118 native speakers of English) through Amazon Mechanical Turk (mean age: 36.4 years; range: 21–72 years). One additional participant submitted a completed HIT, but no corresponding data file was stored to our server for unknown reasons. Our rationale for our sample size was that for these more tightly controlled stimuli, the true condition effect was likely to be smaller than in Experiments 3A-3B. A sample size of approximately 120 participants was chosen to have 80% power to detect a medium effect. Participants were randomly assigned to the High Nameability Condition (n = 58) or to the Low Nameability Condition (n = 61) and were paid $0.60 for completing the task (average completion time: 4.3 min, SD = 1.5).

7.1.2. Stimuli

7.1.2.1. Stimulus selection and norming. The stimuli were constructed from an existing set of shape stimuli known as “tangrams” (Atkinson, Mills, & Smith, 2019). Tangrams are novel shapes constructed by combining triangles to form recognizable objects (e.g., animal-like or human-like forms, see Table 5). We selected 18 items from an existing tangram set and created rotated versions of each stimulus that appeared subjectively less easy to name. Next, we collected norming data for the 18 pairs of tangram items both in their canonical orientation and in their rotated position. Our goal was to select the 6 item pairs with the largest difference in nameability between the canonical and rotated orientation, since we predicted that these items would elicit the largest difference in category learning.

We collected naming ratings through Amazon Mechanical Turk (N = 44). Half of the participants were randomly assigned to name the 18 original tangram items in their canonical orientation and half were randomly assigned to name the 18 rotated items. We computed Simpson's diversity index for each item. As expected, the 18 items in canonical orientation (Simpson's diversity index: M = 0.081, 95% CI = [0.061, 0.101]) were more nameable than the 18 items in their rotated orientation (Simpson's diversity index: M = 0.013, 95% CI = [0.008, 0.018]), $t(34) = 6.85$, $p < .001$. We selected the 6 item pairs that had the largest difference in Simpson diversity (and thus had the largest difference in nameability), after removing one item that was highly similar to one of the selected items (two items in the original 18 were camel-like images). The resulting shapes in the high nameability condition were substantially more nameable (Simpson's diversity index: M = 0.113, 95% CI = [0.079, 0.148]) than their rotated counterparts in the low nameability condition (Simpson's diversity index: M = 0.011, 95% CI = [0.0004, 0.022]), $t(10) = 7.21$, $p < .001$.

7.1.2.2. Category exemplars. The category exemplars were constructed by organizing the 6 item pairs for each condition (high vs. low nameability) into two groups of three to create the prototype category exemplars, similar to Experiment 3A. The high and low nameability items of each pair were yoked such that each category exemplar in the high and low condition was composed of the same items, but either in the (more nameable) canonical orientation or the (less nameable) rotated orientation (see Fig. 10A).

7.1.3. Design & procedure

The procedure was identical to that in Experiment 3A.

7.2. Results

7.2.1. Main analysis

The analytic approach was identical to Experiment 3A. Overall, participants in the High Nameability condition (M = 72.2%, 95% CI = [67.6%, 76.8%]) were more accurate than participants in the Low Nameability condition (M = 64.4%, 95% CI = [60.2%, 68.6%]), b = 0.42, 95% CI = [0.09, 0.74], $z = 2.54$, $p = .01$ (see Fig. 10B). As in Experiments 3A-B, accuracy improved across trials overall, b = 0.07, 95% CI = [0.05, 0.09], $z = 7.03$, $p < .001$, and increased more rapidly in the High Nameability condition than in the Low Nameability condition, b = 0.04, 95% CI = [0.001, 0.07], $z = 2.02$, $p = .04$. There was no evidence for a speed-accuracy tradeoff (average reaction times in High Nameability Condition: M = 1991 ms, 95% CI = [1847 ms, 2134 ms]; Low Nameability Condition: M = 1887 ms, 95% CI = [1745 ms, 2030 ms]).

7.2.2. Low- vs. high-performing participants in Experiment 4

As in previous experiments, we split participants into a “low learner” group and a “high learner” group based on whether participants achieved at least 75% performance on the final block of the experiment (i.e. was correct on 6 of last 8 trials). 46 participants (38.7% of all participants) were < 75% accurate on the final block (“low learners”). Significantly more low learners belonged to the Low Nameability condition than the High Nameability condition (31 of 46;

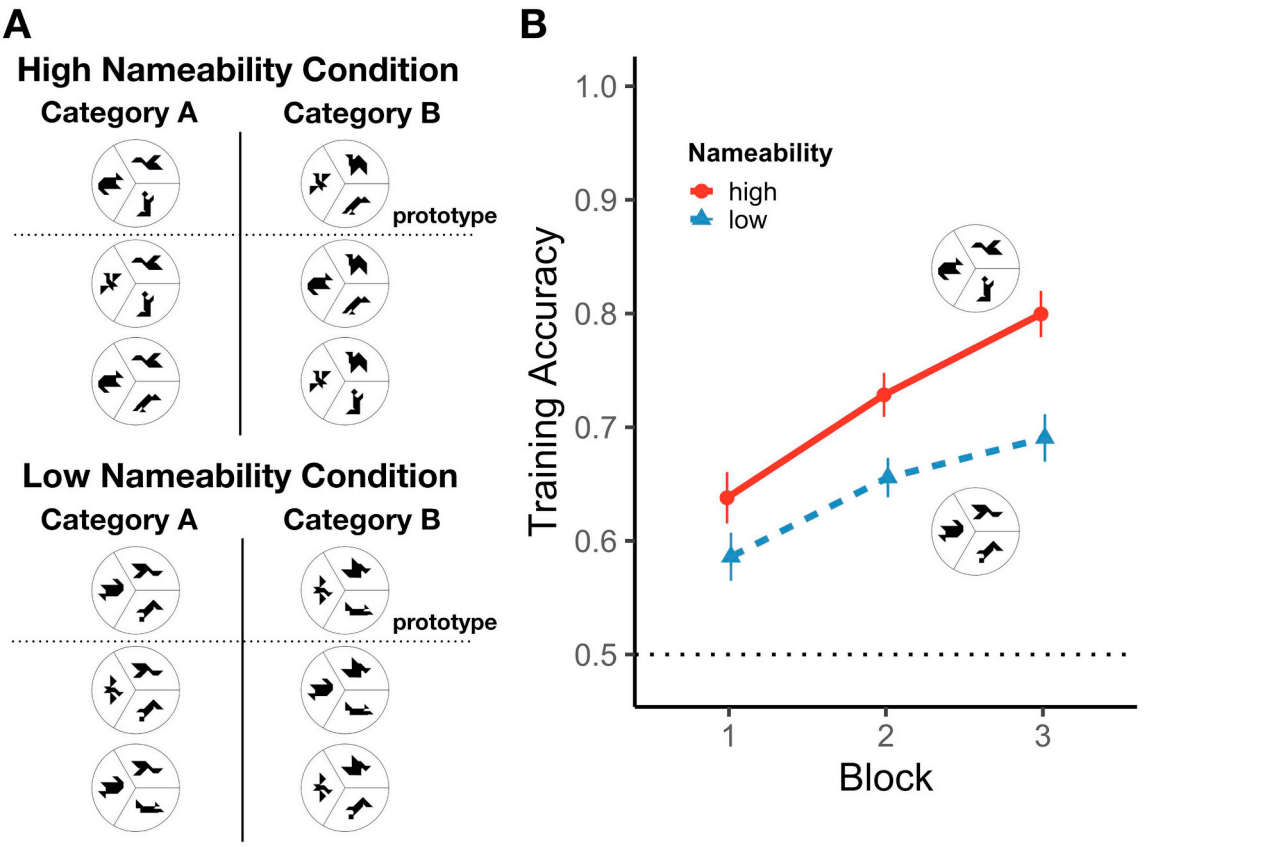

Fig. 10. (A) Stimulus structure in Experiment 4. (B) Accuracy across blocks for Experiment 4. Error bars represent +1/−1 within-subjects SEs (Morey, 2008).

p = .026, exact binomial test), providing additional evidence that the category learning task was significantly more difficult when shape features were rotated into less nameable positions. Unlike in previous experiments, the effect of nameability condition was not significant when reducing the sample to just the high learners (b = 0.25, Wald 95% CI = [−0.16, 0.66], z = 1.19, p = .24), which may be in part due to a lack in power after substantially reducing the total sample.

7.3. Discussion

In Experiment 4, we generalized the results from Experiments 3A-3B using new shapes that more strictly controlled for lower-level perceptual differences, by constructing low nameability shapes that were simple rotations of the shapes from the high nameability condition. This control ensured that high and low nameable shapes were equally matched in any perceptual metric of discriminability or complexity that is orientation-independent. While the effect size in Experiment 4 was smaller than in Experiments 3A-3B, the advantage for more nameable items persisted, suggesting that nameability supports category learning over and above potential low-level perceptual differences. Note that the items in the high nameability condition in Experiment 4 (Average Simpson's Diversity Index = 0.086) were on average less nameable relative to the items in the high nameability condition in Experiments 3A (Average Simpson's Diversity Index = 0.132) and 3B (Average Simpson's Diversity Index = 0.14), while the items in the low nameability condition had similar nameability across all three experiments, which might partially explain both the lower overall performance in the High Nameability condition in Experiment 4 and the smaller effect size. However, this study shows that even small changes in shape orientation that alter nameability can affect how easy it is to learn a new category.

8. General discussion

Categories defined by more nameable color and shape combinations were learned substantially better than categories defined by less nameable features but having an otherwise identical structure. Our work is far from the first to investigate verbal processes in category learning (Ashby et al., 1998; Bruner, Goodnow, & Austin, 1956; Fotiadis & Protopapas, 2014; Minda & Miles, 2010): previous studies have found that categories with more easily verbalized rule structures are easier to learn than categories with rule structures that are more difficult to describe (Ashby & Ell, 2001; Bruner et al., 1956; Kurtz et al., 2013; Shepard et al., 1961). What the present work adds is a demonstration that the likelihood that people successfully use a feature in a category learning task *in the first place* depends on how nameable it is. A compact verbal label may facilitate hypothesis formation: it is easier to pose the hypothesis “it is about redness” than “it is about that pinkish-purplish color”.

Are labels truly the causal force driving the difference in categorization accuracy? We consider four potential confounds as alternate explanations of our results: complexity, familiarity/frequency of exposure, memorability, and meaningfulness.

8.1. Complexity

Although the logical structure of the categories that used easy- and hard-to-name features was exactly the same, a potential concern is that more nameable colors and shapes are more nameable *because* they are somehow cognitively simpler, rather than because labels aid learners in representing them. The burden of this alternative account is to formulate a complexity measure that explains what makes easy-to-name colors and shapes inherently simpler without relying in some form on linguistic measures. Such confounding is common in the literature, such as when the complexity of verbal descriptions of problems is used to

quantify problem complexity (Carpenter et al., 1990) or when simple-to-name properties are assumed to comprise the primitives over which logical complexity is then computed (Feldman, 2003). In the case of shape items, we controlled for shape complexity in Experiments 3A and 3B using an established non-linguistic complexity metric (Feldman & Singh, 2006). Experiment 4 controls by design for any complexity measure that is rotation-independent. This includes the shape skeleton measure of Feldman and Singh (2006), which is insensitive to changes in orientation.

8.2. Familiarity/frequency of exposure

Another worry is that the claimed benefits of nameability reflect differences in familiarity. Familiarity is a purely subjective construct, i.e., to obtain it, we must ask people to judge how familiar something is (Bakhtiar, Nilipour, & Weekes, 2013; Ellis & Morrison, 1998; Liu et al., 2011; Snodgrass & Yuditsky, 1996). What makes something familiar? Subjective ratings of familiarity are closely related to frequency of exposure: people's ratings of the familiarity of word meanings correlate at about .65 with the words' objective print frequencies (Clark & Paivio, 2004; Stadthagen-Gonzalez & Davis, 2006). In evaluating the possibility that it is familiarity rather than nameability that causes the observed differences in categorization accuracy, we therefore reduce (subjective) familiarity to (objective) frequency of exposure and evaluate whether differences in exposure can plausibly explain the learning differences we find in our studies.

Unlike word frequency, which can be easily estimated from printed texts, the frequency with which we are exposed to various colors is harder to estimate. Quantifying the distribution of spectral frequencies is relatively straightforward (Howard & Burnidge, 1994), but quantifying the distribution of *colors* as perceived by human observers is far more complex, because the same spectral input can be seen as entirely different colors depending on, e.g., the surrounding contrast (Lotto & Purves, 2000) and the adapted grey point of the viewer (Webster, 2009). Putting these complexities aside, analyses of color distributions of both natural and urban scenes show that they are dominated by low-saturation colors, which tend to be *less* nameable (Belpaeme & Bleys, 2009; Yendrikhovskij, 2001). Highly nameable focal colors do dominate some types of man-made objects such as children's toys (often in the service of *teaching* children color names, such that color nameability drives the color choices). In other cases, even manufactured colors shy away from highly nameable ones. For example, the family of yellow advertised by Sherman Williams does not seem to include any prototypical yellows.[4] That said, by using colors as stimuli, we relinquish control over the participants' history of exposure. We therefore conducted three studies using novel shapes varying in nameability (Experiments 3A, 3B, and 4). Since participants are unlikely to have previous experience with any of the shapes, any differences in categorization accuracy is likely not due to differences in frequency of prior exposure to the specific items used here.

8.3. Memorability

Another possible confound is memorability. Perhaps the categorization advantage we observed for categories with more nameable features arises from people being better able to remember these features from trial to trial. Nameable colors are (almost by definition) focal colors and focal colors are easier to remember (Bae, Olkkonen, Allred, & Flombaum, 2015). However, finding that focality and memorability are *associated* does not make memorability a confound. It is a confound if it causes differences in categorization accuracy independently of naming (for this reason, we think our experiments using shape features, especially Experiment 4, are not subject to this particular confound).

Rosch, in her original tests of color universals, found that the Dani, a population with a color system quite different from English, showed superior memory for focal colors derived from English speakers (Heider & Olivier, 1972; Rosch Heider, 1972). This finding would seem to suggest that memorability varies independently of naming. Later work, however, found discrepant results for a different population, the Berinmo. Their patterns of memory were best predicted by their patterns of naming rather than by English patterns of naming (Davidoff, Davies, & Roberson, 1999; Roberson, Davies, & Davidoff, 2000). To the extent that there exist associations between nameability and memory, the relationship may thus run from name-based categorization to memory. It is along these lines that Bae et al. (2015) interpret their findings of a close correspondence between memorability and naming, writing:

> What appears, in aggregated responses, as differences in the memorability of different colors is the consequence of a tendency to categorize colors such that some are better examples of a given category than others, and with some as reasonable examples of more than one category. Colors are more accurately and precisely remembered when they are good examples of their respective categories. (p. 760)

That is, differences in color memorability appear to be caused by categorization processes. If, as we claim, color categorization is itself affected by language (see also Forder & Lupyan, 2019 for direct tests of the effects of language on color discrimination), then differences in memorability may be a consequence rather than a cause of differences in nameability. Suggestive evidence comes from a study showing that verbal interference disrupts color memory (Roberson & Davidoff, 2000). A more thorough test of the causal link between naming and memory could involve exposing people to different naming patterns (Özgen & Davies, 2002), and observing the effects of this training on memory for colors that are easier vs. harder to name using the newly learned scheme.

8.4. Meaningfulness

Finally, one might be concerned that some colors or shapes are more nameable because they are more strongly connected to meaningful concepts. While we cannot rule out this possibility, it is not obvious that category learning is necessarily easier when the features are more meaningful. For example, Murphy and Allopenna (1994) found that categories composed of meaningful features such as “Lives alone” and “Has barbed tail” were not learned better than categories composed of relatively meaningless (though still nameable) features such as “+” and “$” when the meaningful features could not be integrated into coherent wholes. In fact, categories that use more meaningful features may sometimes be harder to learn, to the extent that their very meaningfulness leads participants to bring in associated world knowledge that may be irrelevant to the task (Williams, Lombrozo, & Rehder, 2013; Wisniewski & Medin, 1994). Identifying an abstract shape as being dog-like activates knowledge about dogs that may be completely irrelevant to what needs to be learned.

At the same time, making a stimulus meaningful while holding all else constant has been shown to aid categorization in some cases. For example, Lupyan and Spivey were able to speed up visual search for novel stimuli by informing participants that ⊓ and ⊔ were rotated numbers two and five, respectively (Lupyan & Spivey, 2008). Samaha et al. showed that such linguistically ascribed meaning affects even more basic visual processes (Samaha, Boutonnet, Postle, & Lupyan, 2018; see also Rahman & Sommer, 2008). Notice however that in these cases stimuli were made meaningful by using language. Thus, even if it were the case that meaningfulness helps category learning in our experiments, and the easy-to-name stimuli are more meaningful, meaningfulness is only a confound if differences in meaningfulness arise for reasons unrelated to naming. At present, there is no reason to think that

[4] https://www.sherwin-williams.com/homeowners/color/find-and-explore-colors/paint-colors-by-family/family/yellow (last accessed November 6, 2019).

an easy-to-name color, such as a typical blue, is inherently more meaningful than a harder to name color like lavender.

8.5. How does nameability help?

If naming indeed facilitates categorization, by what means does it do so? One possibility is that when learning categories composed of easier-to-name features, these features are more likely to activate their associated labels consistently across trials, making it easier for the learner to track what different category members have in common. In this scenario, the effect of nameability is an *on-line* effect because the difference between more and less nameable features rests on differential recruitment of labels during the task. A second possibility is that previous experience with naming certain features (in our case certain colors and shapes) has led to learning a more categorical representation of the labeled features. In this scenario, the effect of nameability is *off-line*, in that even if linguistic processes were blocked during the task, the nameability effect would still be obtained. Both of these scenarios are expected to make it easier to pose a hypothesis of the form "it is about redness" during learning, but the underlying mechanisms differ. As our present goal was establishing the existence of a causal link between nameability and categorization, our results do little to distinguish between these two scenarios. However, as a preliminary test of the hypothesis that easier-to-name features lead to better category learning because of on-line recruitment of language, we conducted a version of Experiment 2A with a mild form of verbal interference. Methods and results are presented in Supplementary materials S3 and Fig. S6. Verbal interference did not diminish the benefit of nameability, suggesting either that a more taxing form of verbal interference is necessary to disrupt the on-line influence of language on categorization or that the nameability benefit is primarily an off-line phenomenon.

8.6. Implications for categorization and the effects of language

What are the implications of our finding that it is easier to learn rule-based categories if they consist of easier-to-name features? The first is that it offers an alternative metric of category complexity. A category like "red things" may be easy not because it is "inherently easy" (Feldman, 2003), but because we have previously learned a name that coheres its members, or because when the category is used as a feature, its easy-to-access label makes it more likely that people will notice its presence from one time to the next.

Another reason why it matters that ease of category learning depends on nameability is that what is nameable in one language is not necessarily nameable in another. As argued by Evans and Levinson (2009):

> …languages differ enormously in the concepts that they provide ready-coded in grammar and lexicon. Languages may lack words or constructions corresponding to the logical connectives "if" (Guugu Yimithirr) or "or" (Tzeltal), or "blue" or "green" or "hand" or "leg" (Yélî Dnye). There are languages without tense, without aspect, without numerals, or without third-person pronouns […]. Some languages have thousands of verbs; others only have thirty (Schultze-Berndt, 2000). Lack of vocabulary may sometimes merely make expression more cumbersome, but sometimes it effectively limits expressibility, as in the case of languages without numerals (Gordon, 2004). (p. 435)

Our findings suggest that differences in nameability matter not only for cases of inexpressibility (such as attempting to express the meaning *719* in a language lacking number words), but also in cases where differences are ones of mere "cumbersomeness". Our results suggest that when a language makes a feature expressible using a compact term with high name agreement, the feature may be more learnable than when longer verbal expressions are necessary.

One of our stimulus domains – color – has been the subject of decades of research that has sought to discover principles by which continuous color inputs are discretized into lexical categories (e.g., Berlin & Kay, 1969; Kay, Berlin, Maffi, Merrifield, & Cook, 2010; Regier, Kay, & Cook, 2005; Regier, Kay, & Khetarpal, 2007). Although the variability of color categories observed in different languages is highly constrained (Regier et al., 2007; but see Witzel, 2016), *which* colors are most nameable varies considerably from one language to another (Gibson et al., 2017; Majid et al., 2018). In our own analysis of Majid et al.'s data from 20 languages (including 3 signed languages) the correlations between color nameability are surprisingly small: the correlation between English and other tested languages range from 0.56 (English-Cantonese) to near 0 (English-Yurakare: −0.01), to negative (English-Umpila: −0.19). Majid et al.'s data also indicate that highly nameable shapes in English (square, triangle, circle) are not necessarily easy to name in other languages (see also Roberson et al., 2002).

Our claim that nameability is an important predictor of categorization success makes a straightforward prediction: features and relations that are more nameable in a given language will, other things being equal, result in the speakers of the language being better able to use these features for categorization. In combination with the kinds of stimulus controls we applied here (e.g., Experiment 4), such cross-linguistic studies can further confirm – or disconfirm – our claims about the importance of nameability.

9. Conclusion

Our results show that even simple, one-dimensional categories are learned faster when participants have accessible labels to represent category features. We hypothesize that words can act as priors that help learners rapidly form hypotheses about novel categories. These results extend past studies showing that labels lead to faster category learning (Lupyan et al., 2007; Lupyan & Casasanto, 2015) and that language supports our ability to learn abstract concepts such as exact number (Frank, Everett, Fedorenko, & Gibson, 2008; Gordon, 2004) and relational categories (Christie & Gentner, 2014; Gentner, 2016; Gentner, Ozyürek, Gürcanli, & Goldin-Meadow, 2013). It may also help to explain why humans are much more apt at learning rule-based categories compared to other animal species (Smith, Redford, Haas, Coutinho, & Couchman, 2008). A simple word can be a powerful guide to discovering new joints in the world.

Acknowledgements

This material is supported by NSF-PAC 1734260 to GL and NSF-GRFP DGE-1256259 to MZ. We thank Lynn K. Perry for aiding in developing central ideas underlying this study.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.cognition.2019.104135.

References

Alfonso-Reese, L. A., Ashby, F. G., & Brainard, D. H. (2002). What makes a categorization task difficult? *Perception & Psychophysics, 64*(4), 570–583. https://doi.org/10.3758/BF03194727.

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

Ashby, F. G., & Ell, S. W. (2001). The neurobiology of human category learning. *Trends in Cognitive Sciences, 5*(5), 204–210. https://doi.org/10.1016/S1364-6613(00)01624-7.

Atkinson, M., Mills, G. J., & Smith, K. (2019). Social group effects on the emergence of communicative conventions and language complexity. *Journal of Language Evolution, 4*(1), 1–18. https://doi.org/10.1093/jole/lzy010.

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

Bakhtiar, M., Nilipour, R., & Weekes, B. S. (2013). Predictors of timed picture naming in Persian. *Behavior Research Methods, 45*(3), 834–841. https://doi.org/10.3758/s13428-012-0298-6.

Balota, D. A., Yap, M. J., Cortese, M. J., Hutchison, K. A., Kessler, B., Loftis, B., ... Treiman, R. (2007). The English lexicon project. *Behavior Research Methods, 39*(3), 445–459. https://doi.org/10.3758/BF03193014.

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language, 68*(3), 255–278. https://doi.org/10.1016/j.jml.2012.11.001.

Bates, D., & Maechler, M. (2009). lme4: Linear mixed-effects models using S4 classes. Retrieved from https://cran.r-project.org/web/packages/lme4/index.html

Bates, E., D'Amico, S., Jacobsen, T., Székely, A., Andonova, E., Devescovi, A., ... Tzeng, O. (2003). Timed picture naming in seven languages. *Psychonomic Bulletin & Review, 10*(2), 344–380. https://doi.org/10.3758/BF03196494.

Belpaeme, T., & Bleys, J. (2009). The impact of statistical distributions of colours on colour category acquisition. *Journal of Cognitive Science, 10*(1), 1–20. https://doi.org/10.17791/jcs.2009.10.1.1.

Berlin, B., & Kay, P. (1969). *Basic color terms: Their universality and evolution.* Berkeley, CA: University of California.

Boroditsky, L., Fuhrman, O., & McCormick, K. (2011). Do English and Mandarin speakers think about time differently? *Cognition, 118*(1), 123–129. https://doi.org/10.1016/j.cognition.2010.09.010.

Boroditsky, L., & Gaby, A. (2010). Remembrances of times East: Absolute spatial representations of time in an Australian aboriginal community. *Psychological Science, 21*(11), 1635–1639. https://doi.org/10.1177/0956797610386621.

Brodeur, M. B., Dionne-Dostie, E., Montreuil, T., & Lepage, M. (2010). The bank of standardized stimuli (BOSS), a new set of 480 normative photos of objects to be used as visual stimuli in cognitive research. *PLoS One, 5*(5), https://doi.org/10.1371/journal.pone.0010773.

Brodeur, M. B., Guérard, K., & Bouras, M. (2014). Bank of Standardized Stimuli (BOSS) phase ii: 930 new normative photos. *PLoS One, 9*(9), https://doi.org/10.1371/journal.pone.0106953.

Brown, R. W., & Lenneberg, E. H. (1954). A study in language and cognition. *Journal of Abnormal and Social Psychology, 49*(3), 454–462. https://doi.org/10.1037/h0057814.

Bruner, J. S., Goodnow, J. A., & Austin, G. S. (1956). *A study of thinking.* New York: John Wiley & Sons.

Carpenter, P. A., Just, M. A., & Shell, P. (1990). What one intelligence test measures: A theoretical account of the processing in the Raven Progressive Matrices Test. *Psychological Review, 97*(3), 404–431. https://doi.org/10.1037/0033-295X.97.3.404.

Chater, N. (1999). The search for simplicity: A fundamental cognitive principle? *The Quarterly Journal of Experimental Psychology Section A: Human Experimental Psychology, 52*(2), 273–302. https://doi.org/10.1080/713755819.

Christie, S., & Gentner, D. (2014). Language helps children succeed on a classic analogy task. *Cognitive Science, 38*(2), 383–397. https://doi.org/10.1111/cogs.12099.

Clark, J. M., & Paivio, A. (2004). Extensions of the Paivio, Yuille, and Madigan (1968) norms. *Behavior Research Methods, Instruments, & Computers, 36*(3), 371–383. https://doi.org/10.3758/BF03195584.

Couchman, J. J., Coutinho, M. V. C., & Smith, J. D. (2010). Rules and resemblance: Their changing balance in the category learning of humans (Homo sapiens) and monkeys (Macaca mulatta). *Journal of Experimental Psychology. Animal Behavior Processes, 36*(2), 172–183. https://doi.org/10.1037/a0016748.Rules.

Davidoff, J., Davies, I., & Roberson, D. (1999). Colour categories in a stone-age tribe. *Nature, 398*, 203–204. https://doi.org/10.1038/18335.

Davidoff, J., & Roberson, D. (2004). Preserved thematic and impaired taxonomic categorisation: A case study. *Language and Cognitive Processes, 19*(1), 137–174. https://doi.org/10.1080/01690960344000125.

Ellis, A. W., & Morrison, C. M. (1998). Real age-of-acquisition effects in lexical retrieval. *Journal of Experimental Psychology: Learning Memory and Cognition, 24*(2), 515–523. https://doi.org/10.1037/0278-7393.24.2.515.

Enfield, N. J., Majid, A., & van Staden, M. (2006). Cross-linguistic categorisation of the body: Introduction. *Language Sciences, 28*(2–3), 137–147. https://doi.org/10.1016/j.langsci.2005.11.001.

Evans, N., & Levinson, S. C. (2009). The myth of language universals: Language diversity and its importance for cognitive science. *Behavioral and Brain Sciences, 32*(5), 429–448. https://doi.org/10.1017/S0140525X0999094X.

Feldman, J. (2000). Minimization of Boolean complexity in human concept learning. *Nature, 407*(6804), 630–633. https://doi.org/10.1038/35036586.

Feldman, J. (2003). The simplicity principle in human concept learning. *Current Directions in Psychological Science, 12*(6), 227–232. https://doi.org/10.1046/j.0963-7214.2003.01267.x.

Feldman, J. (2006). An algebra of human concept learning. *Journal of Mathematical Psychology, 50*(4), 339–368. https://doi.org/10.1016/j.jmp.2006.03.002.

Feldman, J., & Singh, M. (2006). Bayesian estimation of the shape skeleton. *Proceedings of the National Academy of Sciences, 103*(47), 18014–18019. https://doi.org/10.1073/pnas.0608811103.

Forder, L., & Lupyan, G. (2019). Hearing words changes color perception: Facilitation of color discrimination by verbal and visual cues. *Journal of Experimental Psychology: General, 148*(7), 1105–1123. https://doi.org/10.1037/xge0000560.

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

Frank, M. C., Everett, D. L., Fedorenko, E., & Gibson, E. (2008). Number as a cognitive technology: Evidence from Pirahã language and cognition. *Cognition, 108*(3), 819–824. https://doi.org/10.1016/j.cognition.2008.04.007.

Gentner, D. (2016). Language as cognitive tool kit: How language supports relational thought. *American Psychologist, 71*(8), 650–657. https://doi.org/10.1037/amp0000082.

Gentner, D., Ozyürek, A., Gürcanli, O., & Goldin-Meadow, S. (2013). Spatial language facilitates spatial cognition: Evidence from children who lack language input. *Cognition, 127*(3), 318–330. https://doi.org/10.1016/j.cognition.2013.01.003.

Gibson, E., Futrell, R., Jara-Ettinger, J., Mahowald, K., Bergen, L., Ratnasingam, S., ... Conway, B. R. (2017). Color naming across languages reflects color use. *Proceedings of the National Academy of Sciences of the United States of America, 114*(40), 10785–10790. https://doi.org/10.1073/pnas.1619666114.

Goddard, C., & Wierzbicka, A. (2014). *Words and meanings: Lexical semantics across domains, languages, and cultures.* Oxford: Oxford University Press.

Goldstone, R. L. (2000). Unitization during category learning. *Journal of Experimental Psychology. Human Perception and Performance, 26*(1), 86–112. https://doi.org/10.1037/0096-1523.26.1.86.

Gordon, P. (2004). Numerical cognition without words: Evidence from Amazonia. *Science, 306*(5695), 496–499. https://doi.org/10.1126/science.1094492.

Guest, S., & Van Laar, D. (2002). The effect of name category and discriminability on the search characteristics of colour sets. *Perception, 31*(4), 445–461. https://doi.org/10.1068/p3134.

Hahn, U., Chater, N., & Richardson, L. B. (2003). Similarity as transformation. *Cognition, 87*(1), 1–32. https://doi.org/10.1016/S0010-0277(02)00184-1.

Haun, D. B. M., Rapold, C. J., Janzen, G., & Levinson, S. C. (2011). Plasticity of human spatial cognition: Spatial language and cognition covary across cultures. *Cognition, 119*(1), 70–80. https://doi.org/10.1016/j.cognition.2010.12.009.

Heider, E. R., & Olivier, D. C. (1972). The structure of the color space in naming and memory for two languages. *Cognitive Psychology, 3*(2), 337–354. https://doi.org/10.1016/0010-0285(72)90011-4.

Hjelmquist, E. K. E. (1989). Concept formation in non-verbal categorization tasks in brain-damaged patients with and without aphasia. *Scandinavian Journal of Psychology, 30*(4), 243–254. https://doi.org/10.1111/j.1467-9450.1989.tb01087.x.

Howard, C. M., & Burnidge, J. A. (1994). Colors in natural landscapes. *Journal of the Society for Information Display, 2*(1), 47–55. https://doi.org/10.1889/1.1984908.

Kay, P., Berlin, B., Maffi, L., Merrifield, W. R., & Cook, R. (2010). *The world color survey.* Stanford: CSLI.

Kemp, C., & Regier, T. (2012). Kinship categories across languages reflect general communicative principles. *Science, 336*(6084), 1049–1054. https://doi.org/10.1126/science.1218811.

Kellman, P. J., & Garrigan, P. (2009). Perceptual learning and human expertise. *Physics of Life Reviews, 6*(2), 53–84. https://doi.org/10.1016/j.plrev.2008.12.001.

Koemeda-Lutz, M., Cohen, R., & Meier, E. (1987). Organization of and access to semantic memory in aphasia. *Brain and Language, 30*(2), 321–337. https://doi.org/10.1016/0093-934X(87)90106-4.

Kruschke, J. K. (1992). ALCOVE: An exemplar-based connectionist model of category learning. *Psychological Review, 99*(1), 22–44. https://doi.org/10.1037/0033-295X.99.1.22.

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

de Leeuw, J. R. (2014). jsPsych: A JavaScript library for creating behavioral experiments in a Web browser. *Behavior Research Methods,* 1–12. https://doi.org/10.3758/s13428-014-0458-y.

Levinson, S. C., & Wilkins, D. P. (2006). *Grammars of space.* Cambridge: Cambridge University Press.

Li, P., & Gleitman, L. (2002). Turning the tables: Language and spatial reasoning. *Cognition, 83*, 265–294. https://doi.org/10.1016/S0010-0277(02)00009-4.

Liu, Y., Hao, M., Li, P., & Shu, H. (2011). Timed picture naming norms for Mandarin Chinese. *PLoS One, 6*(1), https://doi.org/10.1371/journal.pone.0016505.

Lotto, R. B., & Purves, D. (2000). An empirical explanation of color contrast. *Proceedings of the National Academy of Sciences, 97*(23), 12834–12839. https://doi.org/10.1073/pnas.210369597.

Love, B. C., Medin, D. L., & Gureckis, T. M. (2004). SUSTAIN: A network model of category learning. *Psychological Review, 111*(2), 309–332. https://doi.org/10.1037/0033-295X.111.2.309.

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

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

Lupyan, G. (2016). The centrality of language in human cognition. *Language Learning, 66*(3), 516–553. https://doi.org/10.1111/lang.12155.

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

Lupyan, G., & Clark, A. (2015). Words and the world: Predictive coding and the language-perception-cognition interface. *Current Directions in Psychological Science, 24*(4), 279–284. https://doi.org/10.1177/0963721415570732.

Lupyan, G., & Mirman, D. (2013). Linking language and categorization: Evidence from aphasia. *Cortex, 49*(5), 1187–1194. https://doi.org/10.1016/j.cortex.2012.06.006.

Lupyan, G., Rakison, D. H., & McClelland, J. L. (2007). Language is not just for talking: Redundant labels facilitate learning of novel categories. *Psychological Science, 18*(12), 1077–1083. https://doi.org/10.1111/j.1467-9280.2007.02028.x.

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

Lupyan, G., & Spivey, M. J. (2010). Making the invisible visible: Verbal but not visual

cues enhance visual detection. *PLoS One, 5*(7), e11452. https://doi.org/10.1371/journal.pone.0011452.
Lupyan, G., & Ward, E. J. (2013). Language can boost otherwise unseen objects into visual awareness. *Proceedings of the National Academy of Sciences, 110*(35), 14196–14201. https://doi.org/10.1073/pnas.1303312110.
Luria, A. R. (1976). *Cognitive development: Its cultural and social foundations.* Oxford, England: Harvard University Press.
Majid, A., Bowerman, M., van Staden, M., & Boster, J. S. (2007). The semantic categories of cutting and breaking events: A crosslinguistic perspective. *Cognitive Linguistics, 18*(2), 133–152. https://doi.org/10.1515/COG.2007.005.
Majid, A., & Burenhult, N. (2014). Odors are expressible in language, as long as you speak the right language. *Cognition, 130*(2), 266–270. https://doi.org/10.1016/j.cognition.2013.11.004.
Majid, A., Roberts, S. G., Cilissen, L., Emmorey, K., Nicodemus, B., O'Grady, L., ... Levinson, S. C. (2018). Differential coding of perception in the world's languages. *Proceedings of the National Academy of Sciences, 115*(45), 11369–11376. https://doi.org/10.1073/pnas.1720419115.
Malt, B. C., Gennari, S., Imai, M., Ameel, E., Saji, N., & Majid, A. (2015). Where are the concepts? What words can and can't reveal. In E. Margolis, & S. Laurence (Eds.). *Concepts: New directions* (pp. 291–326). Cambridge, MA: MIT Press.
Minda, J. P., Desroches, A. S., & Church, B. A. (2008). Learning rule-described and non-rule-described categories: A comparison of children and adults. *Journal of Experimental Psychology. Learning, Memory, and Cognition, 34*(6), 1518–1533. https://doi.org/10.1037/a0013355.
Minda, J. P., & Miles, S. J. (2010). The influence of verbal and nonverbal processing on category learning. In B. H. Ross (Vol. Ed.), *Psychology of learning and motivation-advances in research and theory. Vol. 52. Psychology of learning and motivation-advances in research and theory* (pp. 117–162). Burlington: Academic Press. https://doi.org/10.1016/S0079-7421(10)52003-6.
Mirman, D., Thompson-Schill, S. L., Lupyan, G., & Hamilton, R. (2012). Categorization is modulated by transcranial direct current stimulation over left prefrontal cortex. *Cognition, 124*(1), 36–49. https://doi.org/10.1016/j.cognition.2012.04.002.
Morey, R. D. (2008). Confidence Intervals from normalized data: A correction to Cousineau (2005). *Tutorials in Quantitative Methods for Psychology, 4*(2), 61–64. https://doi.org/10.20982/tqmp.04.2.p061.
Munroe, R. P. (2010). Color survey results. Xkcd https://blog.xkcd.com/2010/05/03/color-survey-results/.
Murdock, G. P. (1970). Kin term patterns and their distribution. *Ethnology, 9*(2), 165–208. https://doi.org/10.2307/3772782.
Murphy, G. L., & Allopenna, P. D. (1994). The locus of knowledge effects in concept learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 20*(4), 904–919. https://doi.org/10.1037/0278-7393.20.4.904.
Nazzi, T., & Gopnik, A. (2001). Linguistic and cognitive abilities in infancy: When does language become a tool for categorization? *Cognition, 80*(3), 11–20. https://doi.org/10.1016/S0010-0277(01)00112-3.
Nosofsky, R. M., Gluck, M. A., Palmeri, T. J., McKinley, S. C., & Glauthier, P. (1994). Comparing modes of rule-based classification learning: A replication and extension of Shepard, Hovland, and Jenkins (1961). *Memory & Cognition, 22*(3), 352–369. https://doi.org/10.3758/BF03200862.
Nosofsky, R. M., & Palmeri, T. J. (1996). Learning to classify integral-dimension stimuli. *Psychonomic Bulletin & Review, 3*(2), 222–226. https://doi.org/10.3758/BF03212422.
Özgen, E., & Davies, I. R. L. (2002). Acquisition of categorical color perception: A perceptual learning approach to the linguistic relativity hypothesis. *Journal of Experimental Psychology: General, 131*(4), 477–493. https://doi.org/10.1037/0096-3445.131.4.477.
Perry, L. K., & Lupyan, G. (2014). The role of language in multi-dimensional categorization: Evidence from transcranial direct current stimulation and exposure to verbal labels. *Brain and Language, 135*, 66–72. https://doi.org/10.1016/j.bandl.2014.05.005.
R Development Core Team (2019). *R: A language and environment for statistical computing.* Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.r-project.org/.
Rabi, R., & Minda, J. P. (2014). Rule-based category learning in children: The role of age and executive functioning. *PLoS One, 9*(1), https://doi.org/10.1371/journal.pone.0085316.
Rahman, R. A., & Sommer, W. (2008). Seeing what we know and understand: How knowledge shapes perception. *Psychonomic Bulletin and Review, 15*(6), 1055–1063. https://doi.org/10.3758/PBR.15.6.1055.
Regier, T., Kay, P., & Cook, R. S. (2005). Focal colors are universal after all. *Proceedings of the National Academy of Sciences, 102*(23), 8386–8391. https://doi.org/10.1073/pnas.0503281102.
Regier, T., Kay, P., & Khetarpal, N. (2007). Color naming reflects optimal partitions of color space. *Proceedings of the National Academy of Sciences, 104*(4), 1436–1441. https://doi.org/10.1073/pnas.0610341104.
Roberson, D., & Davidoff, J. (2000). The categorical perception of colors and facial expressions: The effect of verbal interference. *Memory and Cognition, 28*(6), 977–986. https://doi.org/10.3758/BF03209345.
Roberson, D., Davidoff, J., Davies, I. R. L., & Shapiro, L. R. (2005). Color categories: Evidence for the cultural relativity hypothesis. *Cognitive Psychology, 50*(4), 378–411. https://doi.org/10.1016/j.cogpsych.2004.10.001.
Roberson, D., Davidoff, J., & Shapiro, L. (2002). Squaring the circle: The cultural relativity of "good" shape. *Journal of Cognition and Culture, 2*(1), 29–51. https://doi.org/10.1163/156853702753693299.
Roberson, D., Davies, I., & Davidoff, J. (2000). Color categories are not universal: Replications and new evidence from a stone-age culture. *Journal of Experimental Psychology: General, 129*(3), 369–398. https://doi.org/10.1037/0096-3445.129.3.369.
Rosch Heider, E. (1972). Universals in color naming and memory. *Journal of Experimental Psychology, 93*(1), 10–20. https://doi.org/10.1037/h0032606.
Rossion, B., & Pourtois, G. (2004). Revisiting Snodgrass and Vanderwart's object pictorial set: The role of surface detail in basic-level object recognition. *Perception, 33*(2), 217–236. https://doi.org/10.1068/p5117.
Samaha, J., Boutonnet, B., Postle, B. R., & Lupyan, G. (2018). Effects of meaningfulness on perception: Alpha-band oscillations carry perceptual expectations and influence early visual responses. *Scientific Reports, 8*(1), 1–14. https://doi.org/10.1038/s41598-018-25093-5.
Schyns, P. G., Goldstone, R. L., & Thibaut, J. P. (1998). The development of features in object concepts. *The Behavioral and Brain Sciences, 21*(1), 1–54. https://doi.org/10.1017/S0140525X98000107.
Schultze-Berndt, E. (2000). Simple and complex verbs in Jaminjung: A study of event categorization in an Australian language. *Doctoral dissertation, Radboud University, MPI Series in Psycholinguistics*.
Sharma, G., Wu, W. C., & Daa, E. N. (2005). The CIEDE2000 color-difference formula: Implementation notes, supplementary test data, and mathematical observations. *Color Research and Application, 30*(1), 21–30. https://doi.org/10.1002/col.20070.
Shepard, R. N., Hovland, C. I., & Jenkins, H. M. (1961). Learning and memorization of classifications. *Psychological Monographs: General and Applied, 75*(13), 1–42. https://doi.org/10.1037/h0093825.
Shiffrin, R. M., & Steyvers, M. (1997). A model for recognition memory: REM—Retrieving effectively from memory. *Psychonomic Bulletin & Review, 4*(2), 145–166. https://doi.org/10.3758/BF03209391.
Simpson, E. H. (1949). Measurement of diversity. *Nature, 163*(1943), 688. https://doi.org/10.1038/163688a0.
Slobin, D. I., Ibarretxe-Antunanõ, I., Kopecka, A., & Majid, A. (2014). Manners of human gait: A crosslinguistic event-naming study. *Cognitive Linguistics, 25*(4), 701–741. https://doi.org/10.1515/cog-2014-0061.
Smith, J. D., Redford, J. S., Haas, S. M., Coutinho, M. V., & Couchman, J. J. (2008). The comparative psychology of same-different judgments by humans (Homo sapiens) and monkeys (Macaca mulatta). *Journal of Experimental Psychology: Animal Behavior Processes, 34*(3), 361–374. https://doi.org/10.1037/0097-7403.34.3.361.
Snodgrass, J. G., & Yuditsky, T. (1996). Naming times for the Snodgrass and Vanderwart pictures. *Behavior Research Methods, Instruments, and Computers, 28*(4), 516–536. https://doi.org/10.3758/BF03200540.
Stadthagen-Gonzalez, H., & Davis, C. J. (2006). The Bristol norms for age of acquisition, imageability, and familiarity. *Behavior Research Methods, 38*(4), 598–605. https://doi.org/10.3758/BF03193891.
Steels, L., & Belpaeme, T. (2005). Coordinating perceptually grounded categories through language: A case study for colour. *Behavioral and Brain Sciences, 28*, 469–529. https://doi.org/10.1017/S0140525X05000087.
Székely, A., D'Amico, S., Devescovi, A., Federmeier, K., Herron, D., Iyer, G., ... Bates, E. (2003). Timed picture naming: Extended norms and validation against previous studies. *Behavior Research Methods, Instruments, & Computers: A Journal of the Psychonomic Society, Inc, 35*(4), 621–633. https://doi.org/10.3758/BF03195542.
Vanderplas, J. M., & Garvin, E. A. (1959). The association value of random shapes. *Journal of Experimental Psychology, 57*(3), 147–154. https://doi.org/10.1037/h0048723.
Vigo, R. (2006). A note on the complexity of Boolean concepts. *Journal of Mathematical Psychology, 50*(5), 501–510. https://doi.org/10.1016/j.jmp.2006.05.007.
Waldron, E. M., & Ashby, F. G. (2001). The effects of concurrent task interference on category learning: Evidence for multiple category learning systems. *Psychonomic Bulletin and Review, 8*(1), 168–176. https://doi.org/10.3758/BF03196154.
Webster, M. A. (2009). Calibrating color vision. *Current Biology, 19*(4), R150–R152. https://doi.org/10.1016/j.cub.2008.11.051.
Whorf, B. L. (1956). Language, thought, and reality: Selected writings of Benjamin Lee Whorf. (J. B. Carroll, Ed.), Language, thought, and reality: Selected writings of Benjamin Lee Whorf. Oxford, England: Technology Press of MIT.
Williams, J. J., Lombrozo, T., & Rehder, B. (2013). The hazards of explanation: Overgeneralization in the face of exceptions. *Journal of Experimental Psychology: General, 142*(4), 1006–1014. https://doi.org/10.1037/a0030996.
Wisniewski, E., & Medin, D. L. (1994). On the interaction of theory and data in concept learning. *Cognitive Science, 18*(2), 221–281. https://doi.org/10.1016/0364-0213(94)90002-7.
Witzel, C. (2016). New insights into the evolution of color terms or an effect of saturation? *I-Perception, 7*(5), 1–4. https://doi.org/10.1177/2041669516662040.
Yendrikhovskij, S. N. (2001). Computing color categories from statistics of natural images. *Journal of Imaging Science and Technology, 45*(5), 409–417.