

## Nameability Supports Rule-based Category Learning in Children and Adults

Martin Zettersten<sup>1,2\*</sup>, Catherine Bredemann<sup>3</sup>, Megan Kaul<sup>3</sup>, Kaitlynn Ellis<sup>3</sup>, Haley A. Vlach<sup>3</sup>,  
Heather Kirkorian<sup>4</sup> & Gary Lupyan<sup>1</sup>

*[Accepted manuscript in press at Child Development]*

<sup>1</sup>University of Wisconsin-Madison, Department of Psychology, 1202 W Johnson St, Madison, WI 53706, USA

<sup>2</sup>Princeton University, Department of Psychology, South Dr, Princeton, NJ 08540, USA

<sup>3</sup>University of Wisconsin-Madison, Department of Educational Psychology, 1025 W Johnson St, Madison, WI 53706, USA

<sup>4</sup>University of Wisconsin-Madison, Human Development and Family Studies Department, 1300 Linden Dr, Madison, WI 53706, USA

**Corresponding author:** Martin Zettersten, Department of Psychology, South Dr, Princeton, NJ, 08540, USA, [martincz@princeton.edu](mailto:martincz@princeton.edu)

### Acknowledgements

The data and reproducible analytic code necessary to reproduce the analyses and the materials necessary to attempt to replicate the findings presented in this paper are publicly accessible in a repository on the Open Science Framework (OSF) at <https://osf.io/3y4ck/>. The analyses presented here were not preregistered. This work was supported by the UW2020 Grant to GL, HV, and HK, and NSF-GRFP DGE-1747503 and NIH NICHD F32HD110174 to MZ. Special thanks to the staff at the UW-Madison Learning, Cognition, & Development Lab for their support during data collection and all the families who participated.

Word count: 9,697

### Abstract

The present study tested the hypothesis that verbal labels support category induction by providing compact hypotheses. 97 4-6-year-old children ( $M=63.2$  months; 46 female, 51 male; 77% White, 8% more than one race, 4% Asian, 3% Black; tested 2018) and 90 adults ( $M=20.1$  years; 70 female, 20 male) in the Midwestern United States learned novel categories with features that were easy (e.g., “red”) or difficult (e.g., “mauve”) to name. Adults ( $d = 1.06$ ) and – to a lesser extent – children ( $d = 0.57$ ; final training block) learned categories composed of more nameable features better. Children’s knowledge of difficult-to-name color words predicted their learning for categories with difficult-to-name features. Rule-based category learning may be supported by the emerging ability to form verbal hypotheses.

Word count: 120

*Keywords:* rule learning, color, category learning, language, nameability, concepts

### **Nameability Supports Rule-based Category Learning in Children and Adults**

Learning abstract, rule-based categories is crucial to children's development. Rule-based concepts lay the foundation for reasoning across virtually all aspects of human life, from knowing how to maneuver through the world on a daily basis (e.g., knowing to stop at a red light and to go at a green light) to engaging in abstract thought (e.g., learning that triangles are three-sided objects). However, learning rule-based categories also represents a significant hurdle for children. Although five-year-old children can perform complex tasks that require abstraction, such as drawing recognizable common objects (Long et al., 2018, 2019) or engaging in complex pretend play (Lillard et al., 2013), even formally simple rule-based categories can often prove surprisingly difficult – though not impossible – for them to learn (Mathy et al., 2015; Minda et al., 2008; Rabi & Minda, 2014). In this paper, we explored one potential explanation for why children's ability to learn abstract categories improves across development: their growing ability to generate and deploy verbal labels in the service of discovering novel categories.

#### **Children's Learning of Abstract and Rule-based Categories**

Acquiring abstract concepts – concepts that require extracting general patterns from more concrete or specific motor and perceptual experiences – lays the foundation for cognitive development across many domains. One example is the development of reasoning about relational categories – categories defined by roles or properties connecting multiple entities (Christie & Gentner, 2014; Gentner, 2016). Basic relations such as the concepts of “same” and “different” have long been considered “the very keel and backbone of our thinking” (James, 1890/1950, p. 459), laying the foundation for abstract, combinatorial thought. Relational categories play a central role in reasoning across a wide variety of areas, including spatial cognition (Gentner et al., 2013; Loewenstein & Gentner, 2005) and mathematical reasoning

(Singley & Bunge, 2014). Although recent evidence suggests that young children may have abilities to reason about relations such as same and different in certain contexts (Anderson et al., 2018; Walker & Gopnik, 2014), reasoning about relational categories remains quite difficult for children into their preschool years (Christie & Gentner, 2014; Hochmann et al., 2017).

Past research has also investigated the development of children's ability to reason about categories that require learning a specific, usually abstract, category rule (Deng & Sloutsky, 2016; Kloos & Sloutsky, 2008; Minda & Miles, 2010). These studies often find limited or mixed evidence for rule-based category learning in children. On one hand, children between the ages of 3 and 8 can show adult-like category learning for rule-based categories when stimulus dimensions are transparent (e.g., successfully learning to group stimuli based on size as opposed to color) (Minda et al., 2008). On the other hand, children around the same ages often fail to show robust learning when the stimulus dimensions are more opaque or difficult to identify (Rabi & Minda, 2014), when participants are required to learn category rules that combine information from multiple dimensions (Mathy et al., 2015; Minda et al., 2008), or when they are required to suppress a prepotent response or flexibly shift between category rules (Munakata et al., 2012; Zelazo & Carlson, 2012). These contrasting findings raise a puzzle: why are some categories based on logically simple rules so difficult for children to learn?

When encountering a novel category stimulus, a key task for any learner is identifying the dimensions that determine category membership, while learning to ignore task-irrelevant dimensions. On many theories of categorization and its development, such as COVIS (Ashby et al., 1998), children are thought to automatically represent task-relevant and task-irrelevant category dimensions, because the dimensions are considered perceptually basic (e.g., orientation and spatial frequency) (Huang-Pollock et al., 2011) or because children are explicitly taught the

category dimensions (Visser & Raijmakers, 2012). According to these approaches, the main developmental challenge in learning rule-based categories is that children often fail to focus their attention selectively on a single relevant category dimension while inhibiting irrelevant dimensions (Best et al., 2013; Munakata et al., 2012; Plebanek & Sloutsky, 2017). For example, children will fail to learn a single-dimensional category rule when a category-irrelevant feature has high variability (Huang-Pollock et al., 2011). Conversely, children sometimes outperform adults on tasks that require retaining information about category-irrelevant dimensions (Best et al., 2013; Deng & Sloutsky, 2016). However, the notion that learners are attending to specific category dimensions presupposes that they are successfully representing the dimension at hand. Crucially, not all dimensions are equally easy to represent or reason about (Kurtz et al., 2013), and the ability to represent relevant category dimensions undergoes a substantial learning process (Mash, 2006; Schyns et al., 1998; Schyns & Rodet, 1997). This raises the question of how children come to learn which dimensions to represent in the first place.

### **Verbal Labels as Guides to Category Membership**

One tool that may aid in abstracting and generalizing relevant dimensions of categories is the use of verbal labels (Lupyan & Zettersten, 2021). Past research suggests that language is not simply a means for communicating about categories, but also can aid in constructing categories themselves (Carey, 2011; Gentner, 2016; Lupyan, 2016). Verbal labels influence the ability of children and adults to reason about exact number (Frank et al., 2008; Schneider et al., 2020), relational categories (Christie & Gentner, 2014), and spatial concepts (Casasola et al., 2009; Gentner et al., 2013; Miller et al., 2016). Among adult learners, many studies have noted the correlation between the learnability of a formal category rule and the simplicity with which it can be expressed in language (Kurtz et al., 2013; Shepard et al., 1961). Teaching names for novel

categories that are difficult to verbalize leads to improved category learning (Lupyan et al., 2007; Lupyan & Casasanto, 2015), suggesting a causal role for language. One reason that labels support category learning is that they aid in representing category-relevant dimensions (Lupyan & Zettersten, 2021; Perry & Lupyan, 2014). On this view, labels provide compact, easily accessible representations for reasoning about the dimensions of a novel stimulus.

In a recent study, Zettersten & Lupyan (2020) found that the nameability of category features predicts category learning accuracy, controlling for category rule complexity and perceptual discriminability of category features. Participants were tasked with learning novel categories composed of colors or shapes and defined by simple one-dimensional rules (e.g., “images that have the color red belong to category A”). Despite the simplicity of the underlying rule, participants learned the novel categories more quickly and accurately when they were composed of features that were easy to name than when the features were harder to name (e.g., “images that have that yellow-greenish color belong to category A”). These findings suggest that the degree to which underlying category dimensions are easy to access verbally can substantially impact how quickly and easily adult learners identify novel categories.

Verbal labels, such as words for specific color categories, could also play a role in how children approach novel categories. The availability of verbal labels has been shown to affect children’s success at abstract reasoning tasks. For example, 2-4-year-olds only succeeded in relational match-to-sample tasks when given training with labels that highlighted the relation (e.g., “same”) (Christie & Gentner, 2014). The availability of language describing spatial relations helps preschool children remember spatial locations and encode relative location in spatial reasoning tasks (Gentner et al., 2013; Miller et al., 2016; Simms & Gentner, 2019). A possible explanation for why verbal labels support children’s success on such tasks is that labels

help children focus on, represent, and remember features of the stimuli (i.e., “sameness” or a relative location such as “under the box”) that are difficult to conceptualize or are overridden by more salient perceptual information (Gentner & Christie, 2010; Overkott et al., 2023).

Another commonly diagnostic category feature that undergoes substantial development in children’s knowledge are color categories. Color words are at least partially understood early in development (starting around 18 months of age; Forbes & Plunkett, 2018, 2019; Wagner, Jergens, & Barner, 2018), though children’s knowledge of color terms undergoes gradual, highly variable development over the first several years of life (Wagner et al., 2013; Yurovsky et al., 2015). Even children as old as five years often struggle to name colors that lie outside prototypical shades (Saji et al., 2020). This wide variability in the accessibility and knowledge of color labels provides an opportunity to test the role of verbal labels in children’s rule-based category learning. If children use color labels to guide rule discovery in category learning, children’s ability to learn categories based on color features should vary based on their knowledge of labels for relevant category features.

### **The Present Study**

Why are some logically simple rule-based categories so difficult for children to learn? In the present study, we tested whether language experience plays a role in explaining this developmental puzzle, by investigating whether children would learn novel categories better when they were composed of more nameable features and whether these benefits were similar in magnitude to those previously observed for adults. We presented children (4- to 6-year-olds) and adults with “color wheels” composed of three colors. Each color wheel belonged to one of two categories, as determined by a single color feature (see Figure 1A; Zettersten & Lupyan, 2020). Critically, the colors composing the color wheel were either highly nameable (High Nameability

condition) or more difficult to name (Low Nameability condition). In addition to testing participants' category learning, we also collected information on their vocabulary and the extent to which they named the colors used in the task.

In our investigation, we focused on children between the ages of four to six years of age because (1) past research has found that children at this age have substantial difficulty learning even simple category rules (Minda et al., 2008; Rabi & Minda, 2014) and (2) children at this age have been shown to benefit substantially from the availability of verbal labels (e.g., Christie & Gentner, 2014; Overkott et al., 2023). While we predicted that adults would be more likely than children to learn the underlying category rule, consistent with past category learning literature, the critical question was whether children would benefit from more nameable features that make it easier for them to represent and test verbal hypotheses about category membership. If children can also use their knowledge of verbal labels to guide rule-based category learning, children should show a benefit for learning categories composed of highly nameable dimensions, similar to the benefit of nameability found in adults. Moreover, given that children's knowledge of color terms varies widely in this age range (Wagner et al., 2013; Yurovsky et al., 2015), individuals' knowledge of relevant labels for category features should predict their category learning success. Children with better knowledge of difficult-to-name color features should find it easier to learn the underlying category rule, because having access to labels should make it easier for them to represent and test the discrete category features in the task.

## Methods

### Participants

**Adult Sample.** Ninety students at a large public university in the Upper Midwest United States (mean age = 20.1 years,  $SD = 1.2$ ; range: 18 – 23 years; 70 female, 20 male; 83 L1



English speakers; tested in 2018; demographic information on race and ethnicity was not collected) participated for course credit. Participants were randomly assigned to the High Nameability ( $n = 45$ ) or the Low Nameability ( $n = 45$ ) condition. A target sample of  $N = 90$  was chosen based on a power analysis showing that a sample of this size gave us at least 80% power to detect an effect of  $d = 0.6$  and larger. The effect size  $d = 0.6$  was chosen based on the smallest effect observed in the previous experiments the task was modeled on (Zettersten & Lupyan, 2020) and based on initial pilot data with children that was consistent with an effect of this size. The average completion time for the study was 7.9 minutes ( $SD = 0.68$ ; Round 1 Training Phase:  $M = 3.4$  mins; Round 2 Training Phase:  $M = 2.5$  mins; Generalization Phase:  $M = 0.7$  mins).

**Child Sample.** Ninety-seven children in the Midwestern United States (mean age = 63.2 months,  $SD = 7.0$ ; range: 48 - 81 months; 46 female, 51 male; 77% White, 8% more than one race, 4% Asian, 3% Black, 7% did not disclose; 2% Hispanic or Latino; all L1 English speakers, 7 bilingual; self-reported parental education: 51.5% postgraduate, 30.9% college graduate, 7.2% trade, technical, or vocational training, 5.2% some college, 5.2% no response; household income: \$100,000 or more for 47.4% of families, \$50,000 - \$99,999 for 23.7%, less than \$50,000 for 7.2%, 21.6% preferred not to disclose or did not respond; tested in 2018) were recruited from a preschool database belonging to a child development laboratory at a large public university in the Upper Midwest. Sixteen additional participants were excluded due to technical issues (e.g., a browser issue while running the experiment;  $n = 5$ ), experimenter error (e.g., an error in administering the experiment script;  $n = 5$ ), the child not completing the experiment ( $n = 5$ ), or the child not being fluent in English ( $n = 1$ ). We did not collect information about color blindness, but all participants were highly accurate in identifying the highly nameable colors based on their canonical names (see Color Word Comprehension results). The final sample

includes slightly more participants than the original target sample of 90 participants, because we typically recruit additional participants to ensure that the target sample size is met after exclusions. The average completion time for the study was 12.1 minutes ( $SD = 2.76$ ; Round 1 Training Phase:  $M = 5.2$  mins; Round 2 Training Phase:  $M = 4.0$  mins; Generalization Phase:  $M = 1.1$  mins). Children were given storybooks as compensation for participation. Participants were randomly assigned to the High Nameability ( $n = 49$ ) or the Low Nameability ( $n = 48$ ) condition.

### **Stimuli & Design**

**Stimulus Structure.** Participants were presented with eight circular “color wheel” exemplars, each composed of 3 different colors (see Figure 1A), based on a design used in past category learning studies (Couchman et al., 2010; Zettersten & Lupyan, 2020). One of the colors was always 100% predictive of category membership. For example, in the High Nameability condition, stimuli containing a red segment always belonged to one category, while stimuli containing a brown segment always belonged to the other. The other two color features were correlated with category membership at 75%. Color pairs were tied to specific locations; for instance, the colors in the bottom segment of the circle were either blue or yellow in the High Nameability condition (see Table 1). The position of the 100% predictive color was always the upper-right segment. The stimulus containing the three colors that occurred most frequently with each category was termed the “prototype”. The two other training exemplars in each category differed from the prototype with respect to one of the two 75% predictive colors. During the generalization phase, participants were tested on the two prototype exemplars and two novel items, one belonging to each category. The novel stimuli (not viewed during the training phase) differed from the prototype with respect to both 75% predictive colors and were termed the “novel generalization exemplars” (see below for further information).

**Color Nameability.** The critical manipulation involved the nameability (Guest & Laar, 2002) of the colors comprising each color wheel exemplar. The individual color features were selected based on a large-scale online study in which people were asked to name colors (Munroe, 2010), using a procedure described in detail in a past study of the effects of nameability on category learning using the same color stimuli (Zettersten & Lupyan, 2020). We used Simpson’s diversity index  $D$  (Majid et al., 2018; Simpson, 1949) to quantify nameability. Simpson’s diversity index provides a measure of naming diversity that accounts for both type and frequency of labels generated for a stimulus (Majid et al., 2018). 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 0 to 1, with 0 indicating low nameability (all respondents gave unique labels, i.e.,  $n_i = 1$  for all  $i$ ) and 1 indicating high nameability (all respondents gave the same labels, i.e.,  $i = 1$  and  $n_i = N$ ).

**Color Selection.** The specific color features of the stimuli in the present study matched those used in a past study testing the effect of feature nameability on category learning in adults (Zettersten & Lupyan, 2020: Exp 1B). In this study, 6 colors with high nameability and 6 colors with low nameability were selected such that (a) each color pair was clearly discriminable from the other and (b) the resulting prototypes had approximately equal between-color perceptual discriminability. In the Supplementary Materials (section S1), we describe the methods and metrics used to compute discriminability and the stimulus selection procedure from Zettersten & Lupyan (2020) in further detail.

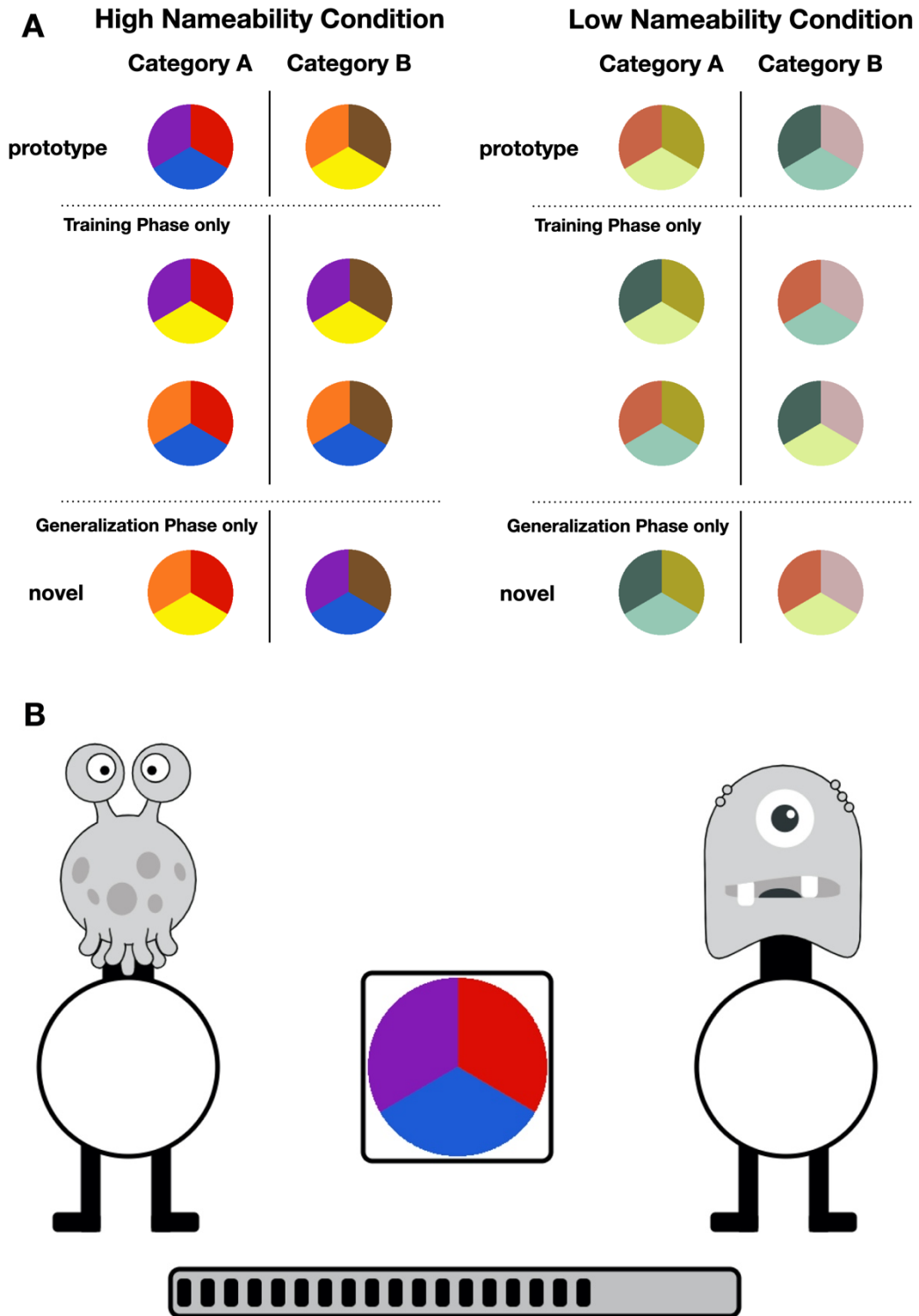














Figure 1. (A) Experimental stimuli and category structure in the training phase. (B) Task design. Participants were asked which alien would like to eat the “snack” (color wheel) and were provided feedback after their selection. The progress bar depicts the number of “snacks” (trials) remaining.

**Table 1***Overview of the color feature set*

RGB	Color	Assigned Name	Modal Name	Nameability	Modal Agreement	Simpson's Diversity	Role
(30, 90, 210)		blue	blue	high	80.3%	.671	75% predictive
(250, 120, 30)		orange	orange	high	85.1%	.733	75% predictive
(220, 20, 0)		red	red	high	82.7%	.697	100% predictive
(250, 240, 0)		yellow	yellow	high	81.7%	.664	75% predictive
(120, 80, 40)		brown	brown	high	81.8%	.648	100% predictive
(130, 30, 180)		purple	purple	high	82.1%	.672	75% predictive
(170,160,40)		chartreuse	mustard	low	6.9%	.056	100% predictive
(200, 170, 170)		mauve	grey	low	6.8%	.054	100% predictive
(200, 100, 70)		sienna	brown	low	8.7%	.051	75% predictive
(70, 100, 90)		teal	grey green	low	9.8%	.128	75% predictive
(220, 240, 150)		honeydew	pale green	low	5.3%	.079	75% predictive
(150, 200, 180)		turquoise	green	low	6.0%	.084	75% predictive

*Note.* Modal names and nameability values (modal agreement and Simpson's diversity index) were computed based on a large-scale online color naming study (Munroe, 2010).

**Color Norming Task.** To ensure that high and low nameability colors were similarly perceptually discriminable, Zettersten and Lupyan (2020) used CIE-LAB distances between color pairs to quantify discriminability during feature selection, and also collected behavioral discriminability norms from adult participants for the color features used in the current experiment. The behavioral norming data found broadly comparable reaction times for high and low nameability colors in a speeded same-different judgment task (though high nameability colors were discriminated slightly faster on some trials). However, it was not clear that these norms would be representative of children’s perceptual discriminability judgments. To obtain a behavioral measure of the discriminability of the color features among children in the current age range, we conducted a norming study in which we asked a separate sample of 3-6-year-old children, as well as a separate sample of adults, to make judgments about color pairs in a speeded match-to-sample task (Zettersten et al., 2020). 40 3-6-year-old children in the Midwestern United States (mean age = 56.4 months,  $SD = 6.5$ ; range: 45 - 69 months; 16 female, 22 male, 2 not reported; 87.5% White, 5% more than one race, 5% Asian, 2.5% did not disclose; tested in 2023) participated in the task. Data from 5 additional children were excluded due to random responding (3), for being outside of the targeted age range (1), or for being diagnosed with a developmental disorder (1). We also collected data from a separate sample of 50 adult students at a large public university in the Upper Midwest United States (mean age = 19.8 years,  $SD = 1.2$ ; 48 female, 2 male) using the same task, to obtain a set of perceptual discriminability norms for both children and adults collected under similar conditions.

The task was conducted on a tablet in a quiet room. In the task, one round color swatch appeared at the top of the screen (the standard; see Figure S1 in the Supplementary Materials). Two round color swatches appeared below the standard, one the same color as the standard (the

target) and one a different color from the standard (the foil). Participants were asked to find which of the two lower pictures matched the top picture as quickly and accurately as possible. The color stimuli were the 6 high nameability and the 6 low nameability colors from the main experiment. High nameability colors were paired only with high nameability colors, and low nameability colors were paired only with low nameability colors, because the manipulation of color nameability in the main experiment was between-subjects (i.e., participants in the main experiment only ever saw high nameability colors or low nameability colors, but not both). Each color pair was tested twice per participant, resulting in 30 high nameability trials and 30 low nameability trials per participant.

Overall, there was no evidence for differences in discriminability between low- and high-nameability colors (for a walkthrough of all analyses, see: <https://rpubs.com/zcm/color-rule-kid-norming>). Accuracy was high across the board both for children (High Nameability Colors: 96.2%, 95% CI = [94.4%, 98.0%]; Low Nameability Colors: 96.9%, 95% CI = [95.7%, 98.0%]) and for adults (High Nameability Colors: 99.7%, 95% CI = [99.3%, 100%]; Low Nameability Colors: 99.7%, 95% CI = [99.1%, 100%]). Average reaction times were similar between high nameability and low nameability colors both for children (High:  $M = 1998$  ms, 95% CI = [1924 ms, 2072 ms]; Low:  $M = 2026$  ms, 95% CI = [1981 ms, 2071 ms],  $t(8.25) = .43$ , null model somewhat favored over the alternative model  $BF_{01} = 2.27$ ) and for adults (High:  $M = 697$  ms, 95% CI = [684 ms, 711 ms]; Low:  $M = 708$  ms, 95% CI = [695 ms, 721 ms],  $t(10) = -1.44$ , null model slightly favored over the alternative model  $BF_{01} = 1.69$ ). We also fit a linear mixed-effects model predicting children's trial-by-trial reaction times from the interaction between condition and age, including by-participant random effects for subject (intercept and condition slope) and color pair. While reaction times decreased with age,  $b = -25.2$ ,  $t(37.79) = -2.24$ ,  $p = .03$ , there

was no significant interaction between age and color nameability,  $b = -1.7$ ,  $t(1223) = -0.27$ ,  $p = .79$ , i.e., there was no evidence of a change in the (lack of) nameability effect across age. At the same time, both children's ( $r = -.47$ ,  $p = .01$ ) and adults' ( $r = -.39$ ,  $p = .03$ ) average reaction times for color pairs were correlated with  $\Delta E_{2000}$  distances between colors.

## Procedure

**Stimulus Presentation.** The stimuli were presented in a web-browser on a Samsung tablet computer [Samsung Galaxy Tab S3 with screen dimensions of 23.6 x 17.0 cm]. The task was coded using the jsPsych library (de Leeuw, 2015). The experiment code and stimuli are available on the project's OSF page (<https://osf.io/3y4ck/>).

**Warm-up Phase.** The task was administered in exactly the same manner with child and adult participants. A trained researcher guided participants through the web-based task, providing short instructions on how to play the game and prompting responses as necessary. The experiment began with a short warm-up phase to familiarize participants with the structure of the task. During the warm-up phase, cartoon images of a cat and a dog appeared on either side of the screen. Next, 4 images of two types of “snacks” appeared one-by-one in the center of the screen. The two snack types were images of bones (the “snacks” that the dog character preferred) and fish (the “snacks” that the cat character preferred). Participants were instructed to “touch the tummy of the animal that you think likes to eat the snack in the middle.” Participants received both auditory and visual feedback after each trial indicating whether a response was correct or incorrect. After correct responses, a positive “ta-da” sound was played while the character jumped up and down. After incorrect responses, a short “buzz” sound played while the stimulus moved back to the central location, and children were then instructed to “feed the snack” to an



animal again. All images, including the dog and cat, were presented in grayscale to ensure that participants were not biased to any particular color during the familiarization phase.

**Training Phase.** Next, participants proceeded to the training phase. Two alien characters appeared on either side of the screen (see Figure 1B), together with a progress bar that allowed participants to track the number of remaining trials. The researcher then explained the task to participants using the following script:

In this game, you're going to meet two different aliens who like different kinds of alien snack. This alien [pointing to alien on the left] is a Modi. Modis like one kind of alien snack. This alien [pointing to alien on the right] is a Gazzer. Gazzers like a different kind of alien snack. Now, you have to figure out what kind of alien snack Modis like and what kind of alien snack Gazzers like. You're going to see two different kinds of alien snack; one kind of snack that Modis like and one kind of snack that Gazzers like. Each alien eats only one of the two snacks, and it's your job to learn which snack to feed each alien. When you see an alien snack in the middle, touch the tummy of the alien that you think likes to eat that kind of snack. If the alien likes that kind of snack, then it will jump up and down. If the alien doesn't like that kind of snack, then the snack will go back to the middle and you get to decide where it goes again.

Participants subsequently sorted the color wheels one-by-one to one of the two alien characters by tapping the "stomach" of the alien character (Figure 1B). Participants completed two rounds of 24 training trials. The 24 training trials in each round were grouped into three blocks of 8 training trials each. On each block, participants sorted the prototype exemplar (the top image in Figure 1A) of each category twice, and the remaining two training exemplars of each category

once. The order of the stimuli was randomized within each block. Participants received immediate feedback as in the warm-up phase. After correct responses, in addition to positive auditory feedback and the target character jumping up and down, the target alien's body also changed colors to match the category stimulus features, cycling through each color feature in turn (i.e., if the stimulus was composed of the color features brown, yellow, and orange, the alien's body changed to brown, yellow, and orange as it jumped up and down). Trials were repeated until participants responded correctly, to ensure that children understood the category learning task and to incentivize correct responding. Only participants' first response on each trial was included in subsequent analyses (i.e., repeated responses after incorrect choices were excluded). Correct locations (left or right) for each stimulus were counterbalanced across participants.

After completing the first round of 24 training trials, participants were given a short break and the original instructions were repeated. Participants were not given any indication that the game was a repetition of the task they had just completed. Instead, the task was presented to participants as if they were starting a new game. They then began a second round of experimental trials. The second round proceeded exactly as the first experimental round, with the same experimental design and procedure. We repeated the training in a second round to help ensure that children received enough experience to be able to induce a category rule, since previous research suggests that children often struggle to learn rule-based categories (Rabi et al., 2015).

**Generalization Phase.** After the Training Phase, participants completed a short Generalization Phase, consisting of 8 trials. The goal of the Generalization Phase was to gather exploratory information about differences in the strategies that participants used to solve the

category learning task. However, we did not predict *a priori* for there to be overall differences in generalization accuracy between the high and low nameability condition, due to previously observed patterns of variability in adults' categorization strategies in both the high and low nameability conditions (see S5 in the Supplementary Materials for further discussion).

The generalization trials included four items, presented twice each in random order: the (previously seen) two prototypes of each category and two novel generalization items not seen during training (Figure 1A). The two novel generalization items differed on both of the 75% predictive color features from the prototype (i.e., only the 100% predictive color was shared with other category members). The novel items were designed to explore whether participants learned a single-feature rule as opposed to a multiple-feature category rule, following the design developed in Couchman et al. (2010). Note that the “correct” rule remains ambiguous from the perspective of the learner: a rule based only on the two 100% predictive colors and a strategy that uses a combination of color features (e.g., a two-out-of-three rule such as “the item belongs to category A if at least two out of the three colors red, blue, and purple are present”) are equally predictive of category membership during the category learning phase. Participants' sorting behavior on the novel items can reveal if a learner is using a consistent strategy and, if so, disambiguate what strategy they are using: if a learner consistently uses only the 100% predictive category features (i.e., in the high nameability condition, the colors red and brown), they will consistently sort the novel items “correctly”, i.e. according to the single-dimensional category rule. However, if they consistently use combinations of category features, they will tend to sort the novel items “incorrectly” (e.g., the two-out-of-three rule above would systematically lead to the “incorrect” classification of the novel generalization exemplars). If participants do not consistently sort the novel items correctly or incorrectly, this suggests that they may not have

learned one single rule or strategy for determining category membership. The procedure for the generalization trials mirrored that for training trials, with two key differences: the generalization trials had no accuracy feedback and trials did not repeat after an incorrect response.

**Vocabulary and Color Word Knowledge Test.** At the end of the category learning task, we presented participants with a series of short tests to assess their verbal knowledge and, in particular, their knowledge of verbal labels for the color features used in the experiment. Immediately after completing the test trials, participants completed a color naming test, a color word comprehension test, and a 12-item general vocabulary test. The color naming task was always presented first. The order of the two subsequent comprehension tasks was counterbalanced across participants.

*Color Naming.* In the color naming task, participants were asked to name the 6 high-nameability colors (displayed on the same tablet computer screen) followed by the 6 low-nameability colors (displayed on a subsequent screen). On a given naming trial, the experimenter pointed to each color on the screen in turn and prompted the participant to name the color, while recording verbal responses. We began the naming task with the high nameability colors because we expected children would have considerable difficulty with the low nameability colors and might become confused if asked to name these colors first. By always presenting the (easier) high nameability colors first, we hoped to reduce the likelihood that children would fail to understand the task.

*Color Word Comprehension.* Participants' comprehension of words for each color was assessed using a 6-alternative forced-choice task. For each color, participants were shown all six colors belonging to a given nameability condition, presented side-by-side in a 2-by-3 array on the screen (as in the color naming task). Participants were then prompted to select a given color

with the question “Can you point to [color name]?”. Participants’ responses were recorded as correct or incorrect by the experimenter. The position of each color on the screen was randomized from trial to trial. The six high nameability colors were tested first in a fixed (random) order, followed by the six low nameability colors in a fixed (random) order.

For the high nameability colors, the six color words tested corresponded to the six names that are typically used by English speakers when referring to these colors (*blue, orange, yellow, red, brown, purple*; see Table 1). For the low nameability colors, the most appropriate color name was — by definition — more difficult to determine. For example, the modal names generated by participants in the large-scale norming survey (Munroe, 2010) are often poor descriptors of their respective colors (e.g., “grey” was the most frequent name for RGB=(200, 170, 170)). To select target labels for each of the six low nameability colors, we used the RGB values of the selected colors to search color databases across the internet and selected high-frequency labels for the exact or similar RGB value. Labels were rejected that implied a semantic association with a familiar object (e.g., “olive”) or that contained a high-nameability color within the label (e.g., “light green”).

*Short Vocabulary Measure.* In order to obtain a brief measure of children’s vocabulary, we selected 12 items from the Peabody Picture Vocabulary Test (PPVT-IV) (Dunn & Dunn, 2007). The items were selected to range from easy items (e.g., *cookie*) to more difficult items (e.g., *mammal*) by taking two trials at random from each age band of the PPVT. The final PPVT-IV test items were (in order) *cookie, belt, fence, farm, calendar, dentist, axe, timer, athlete, hydrant, tusk, mammal*.

## Results

The data and scripts documenting the data analysis for all experiments are openly available on the Open Science Framework (<https://osf.io/3y4ck/>). A walkthrough of all analyses reported in the manuscript, including additional modeling information and analyses reported in the Supplementary Materials, is accessible through a web browser at the following link:

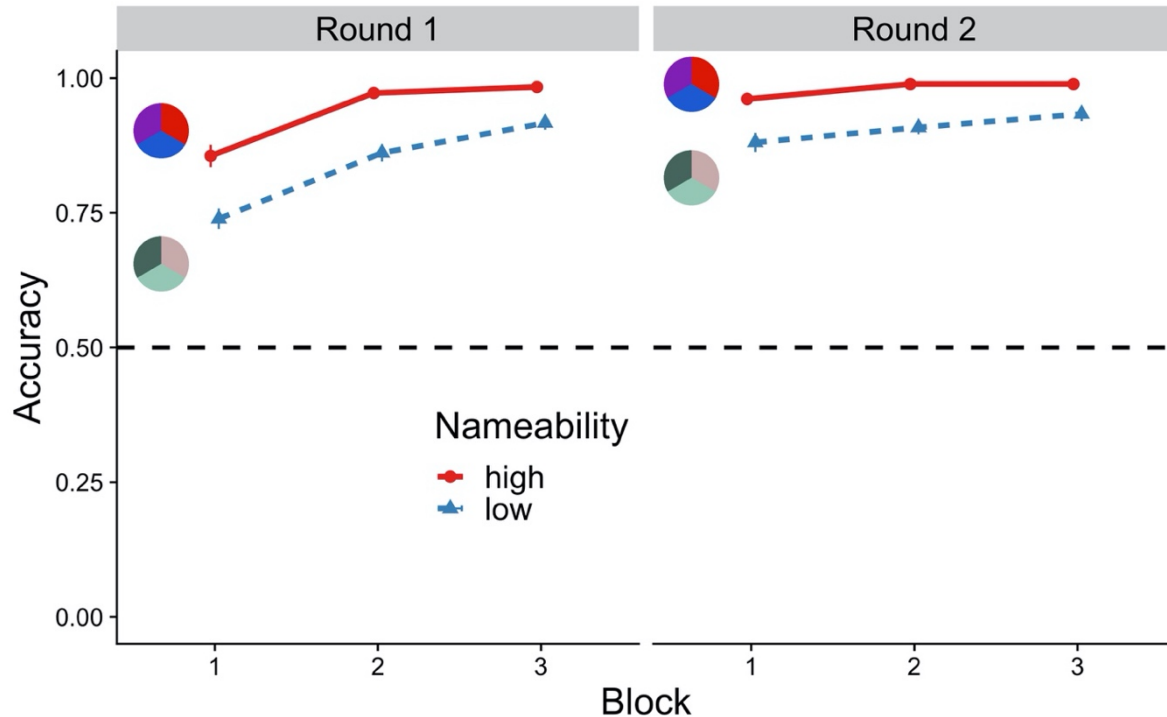
<https://rpubs.com/zcm/color-rule-kid>.

### Category Learning: Training Phase

**Adults.** In our main confirmatory analyses investigating the effect of nameability on category learning, we predicted participants' trial-by-trial accuracy on training trials from Condition (centered; Low Nameability = -0.5, High Nameability = 0.5), Block Number (centered) and Experiment Round (centered), and all interactions between the three predictors in a logistic mixed-effects model (Baayen et al., 2008; Jaeger, 2008). We used the lme4 package version 1.1-31 in R (version 4.2.2) to fit all models (Bates et al., 2015; R Development Core Team, 2022). We fit the model with the maximal by-subject random effects structure, including a by-subject intercept and a by-subject random slopes for Block Number, Experiment Round, and their interaction (Barr et al., 2013).

Table 2 summarizes the coefficients estimated by the model. We highlight four main results. First, participants showed high overall accuracy and performed well above chance in the experiment, as indicated by the intercept term being significantly greater than chance level,  $b = 3.58$ , 95% Wald CI = [3.14, 4.02],  $z = 15.88$ ,  $p < .001$  (chance = 0.5; note that  $\text{logit}(0.5) = 0$ , hence no offset term is needed to test against chance level). Second, critically, participants in the High Nameability condition ( $M = 95.8\%$ , 95% CI = [94.7%, 97.0%]) were more accurate than participants in the Low Nameability Condition ( $M = 87.3\%$ , 95% CI = [84.1%, 90.5%]),  $b =$

1.68, 95% Wald CI = [0.96, 2.39],  $z = 4.61$ ,  $p < .001$  (Figure 2). Third, participants' accuracy increased both across blocks ( $b = 1.26$ , 95% Wald CI = [0.87, 1.66],  $z = 6.23$ ,  $p < .001$ ) and from round 1 to round 2 ( $b = 1.26$ , 95% Wald CI = [.63, 1.88],  $z = 3.94$ ,  $p < .001$ ), providing further evidence that participants learned the categories as the experiment unfolded. Finally, there was a condition-by-block-number interaction, suggesting that participants' accuracy increased more rapidly in the High Nameability condition than in the Low Nameability condition,  $b = 0.56$ , 95% Wald CI = [.03, 1.08],  $z = 2.08$ ,  $p = .038$ . No other interactions were significant.



*Figure 2.* Adults' performance on the category learning task in the High Nameability condition (solid line) and Low Nameability condition (dashed line) during the two rounds of training. Horizontal dashed line indicates chance-level responding. Error bars represent  $\pm 1$  SE of the within-subject corrected mean (Morey, 2008).

**Table 2.**

*Estimates for the logistic mixed-effects model predicting training accuracy for adults*

<b>Coefficient</b>	<b>Estimate</b>	<b>SE</b>	<b>z</b>	<b>p</b>
<b>Intercept</b>	3.58	0.23	15.88	< .001
<b>Condition</b>	1.68	0.36	4.61	< .001
<b>Block Number</b>	1.26	0.20	6.23	< .001
<b>Round</b>	1.26	0.32	3.94	< .001
<b>Condition * Block Number</b>	0.56	0.27	2.08	0.0378
<b>Condition * Round</b>	0.37	0.40	0.92	0.3565
<b>Block Number * Round</b>	-0.23	0.36	-0.64	0.5207
<b>Condition * Block Number * Round</b>	-0.18	0.45	-0.40	0.6932

**Children.** We conducted the same confirmatory test to investigate category learning during the training phase for children (see Table 3). Children's performance was significantly above chance,  $b = .94$ , 95% Wald CI = [.76, 1.11],  $z = 10.40$ ,  $p < .001$ . As with adult participants, children's accuracy improved across blocks ( $b = .12$ , 95% Wald CI = [.03, .22],  $z = 2.55$ ,  $p = .011$ ) and from round 1 to round 2 ( $b = .46$ , 95% Wald CI = [.27, .65],  $z = 4.70$ ,  $p < .001$ ), providing further evidence that children were learning the categories as the experiment progressed.

Performance of children in the High Nameability condition ( $M = 71.0\%$ , 95% CI = [67.2%, 74.8%]) was not significantly more accurate overall than children in the Low Nameability Condition ( $M = 65.9\%$ , 95% CI = [61.6%, 70.3%]),  $b = 0.28$ , 95% Wald CI = [-0.07, 0.63],  $z = 1.56$ ,  $p = .118$  (Figure 3). However, there was a significant condition-by-block-number interaction, suggesting that participants in the High Nameability condition showed faster category learning,  $b = 0.23$ , 95% Wald CI = [0.06, 0.41],  $z = 2.58$ ,  $p = .010$ . Follow-up analyses suggested that this difference was due to participants in the High Nameability condition achieving higher accuracy compared to the Low Nameability condition in the final block (block



3) of round 1 (High:  $M = 74.2\%$ , 95% CI = [69.9%, 78.6%]; Low:  $M = 63.3\%$ , 95% CI = [58.6%, 67.9%];  $b = 0.55$ , 95% Wald CI = [0.15, 0.95],  $z = 2.69$ ,  $p = .007$ ) and round 2 (High:  $M = 75.8\%$ , 95% CI = [71.2%, 80.3%]; Low:  $M = 66.1\%$ , 95% CI = [62.0%, 70.3%];  $b = 0.53$ , 95% Wald CI = [0.07, 0.98],  $z = 2.29$ ,  $p = .022$ ). There was also a significant interaction between block number and round, suggesting that children's learning increased more slowly in round 2 compared to round 1 (i.e., children's learning plateaued),  $b = -0.26$ , 95% Wald CI = [-0.44, -0.07],  $z = -2.70$ ,  $p = .007$ .

**Table 3**

*Estimates for the logistic mixed-effects model predicting training accuracy for children*

<b>Coefficient</b>	<b>Estimate</b>	<b>SE</b>	<b>z</b>	<b>p</b>
<b>Intercept</b>	0.94	0.09	10.40	<0.001
<b>Condition</b>	0.28	0.18	1.56	0.1181
<b>Block Number</b>	0.12	0.05	2.55	0.0109
<b>Round</b>	0.46	0.10	4.70	<0.001
<b>Condition * Block Number</b>	0.23	0.09	2.58	0.0099
<b>Condition * Round</b>	0.07	0.18	0.36	0.7210
<b>Block Number * Round</b>	-0.26	0.09	-2.70	0.0069
<b>Condition * Block Number * Round</b>	-0.12	0.17	-0.68	0.4956

To explore whether these effects were moderated by child age, we fit the same model while adding age (centered), as well as its interaction with all other predictors. There was a significant main effect of Age on children's category learning accuracy,  $b = 0.04$ ,  $z = 3.34$ ,  $p < .001$ . However, age did not moderate the effect of any of the model's predictors, and all of the patterns of significance, including the key Condition by Block Number interaction, remained identical when controlling for age (see Supplementary Materials, S3.2, for details).

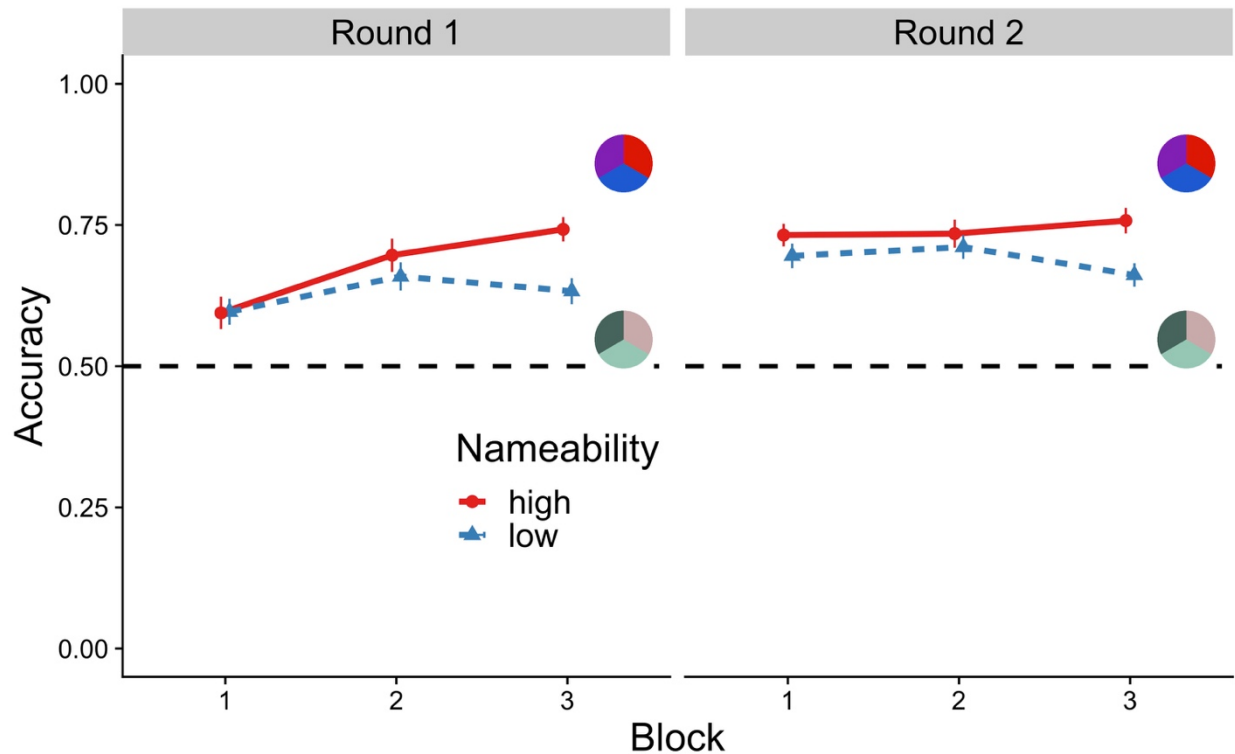


Figure 3. Children's performance on the category learning task during the two rounds of training. Dashed line indicates chance level responding. Error bars represent  $\pm 1$  SE of the within-subject corrected mean (Morey, 2008).

**Comparing Children and Adults.** We tested for differences between children and adults by fitting the same model as that described above across the entire dataset, while including age group as a predictor along with all interactions with the other three predictors: condition, block number and experiment round. We summarize the main results below by focusing on the effects of age group and its interaction with other predictors (see Supplementary Materials, section S3.1 for full model results).

Adults were more accurate overall than children,  $b = 2.32$ , 95% Wald CI = [1.98, 2.67],  $z = 13.18$ ,  $p < .001$ , and their accuracy increased more rapidly across both blocks ( $b = 0.78$ , 95%

Wald CI = [0.54, 1.03],  $z = 6.37$ ,  $p < .001$ ) and rounds ( $b = .68$ , 95% Wald CI = [0.24, 1.11],  $z = 3.04$ ,  $p = .002$ ). Crucially, the overall effect of condition was greater for adults compared to children,  $b = 1.28$ , 95% Wald CI = [0.61, 1.96],  $z = 3.73$ ,  $p < .001$ . However, the condition-by-block-number interaction (significant for both adult and child participants considered separately, see above) did not differ significantly between children and adults,  $b = 0.28$ , 95% Wald CI = [-0.18, 0.74],  $z = 1.19$ ,  $p = .24$ , indicating no evidence for a difference between children and adults in how the accuracy increase across blocks depended on condition.

### **Category Learning: Generalization Phase**

**Adults.** We additionally explored the degree to which nameability influenced participants' performance in the Generalization Phase. Participants were near ceiling in categorizing the prototype stimulus in both nameability conditions (High:  $M = 100\%$ ; Low:  $M = 97.8\%$ , 95% CI = [95.1%, 100%]). Participants sorted the novel items in accordance with a single-color feature rule at similar levels in both conditions (High:  $M = 69.4\%$ , 95% CI = [56.3%, 82.6%]; Low:  $M = 74.4\%$ , 95% CI = [63.0%, 85.9%]). To test for condition differences, we fit a logistic mixed-effects model predicting trial-by-trial accuracy from condition (centered) while controlling for stimulus type (centered). We included a by-subject random intercept and a by-subject random slope for stimulus type. There was no significant difference between conditions,  $b = .15$ , 95% Wald CI = [-2.44, 2.75],  $z = 0.12$ ,  $p = .91$ .

**Children.** Participants categorized the prototype stimulus in both nameability conditions at similar rates (High:  $M = 69.4\%$ , 95% CI = [60.8%, 78.0%]; Low:  $M = 65.6\%$ , 95% CI = [56.8%, 74.4%]; Figure S7B). Participants also sorted the novel generalization exemplars in line with a single-color feature rule at a similar rate in both conditions (High:  $M = 62.2\%$ , 95% CI =

[52.1%, 72.4%]; Low:  $M = 57.8%$ , 95% CI = [48.7%, 66.9%]). There was no significant effect of condition,  $b = .24$ , 95% Wald CI = [-.24, .71],  $z = 0.97$ ,  $p = .33$ .

**Comparing Children and Adults.** Finally, we tested for differences between children and adults by fitting a logistic mixed-effects model predicting trial-by-trial accuracy from condition, stimulus type, age group (centered; children vs. adults) and the 2-way interactions between condition and age group as well as stimulus type and age group. We included a by-subject random intercept and a by-subject random slope for stimulus type. There was a significant effect of age group, revealing that adults were more accurate than child participants,  $b = 2.84$ , 95% Wald CI = [2.07, 3.61],  $z = 7.24$ ,  $p < .001$ . Participants also performed better on the prototype stimuli than on the novel stimuli overall,  $b = 1.55$ , 95% Wald CI = [0.70, 2.41],  $z = 3.55$ ,  $p < .001$ . However, this effect was strongly moderated by age, such that the difference between prototype and novel stimulus accuracy was far greater for adult participants compared to child participants,  $b = 2.74$ , 95% Wald CI = [1.18, 4.30],  $z = 3.44$ ,  $p < .001$ . Neither the effect of condition nor the condition-by-age group interaction term was significant ( $ps > .39$ ; see S5 in the Supplementary Materials for additional analyses).

### **Color Word and Vocabulary Knowledge**

**Color Naming.** As expected, colors from the High Nameability condition were easier to name as measured by Simpson's diversity index of naming responses both among adults (High Nameability colors:  $M = 1.00$ ; Low Nameability colors:  $M = 0.20$ , 95% CI = [0.12, 0.28];  $t(10) = 25.21$ ,  $p < .001$ ) and among children (High Nameability colors:  $M = 0.96$ ; Low Nameability colors:  $M = 0.24$ , 95% CI = [0.12, 0.35];  $t(10) = 14.59$ ,  $p < .001$ ; see Table S5 in the Supplementary Materials). Nameability did not differ between children and adults, either for highly nameable colors ( $t(5) = 1.49$ ,  $p = .20$ ) or for more difficult-to-name colors ( $t(5) = -1.17$ ,  $p$

= .29). The low nameability colors with the highest (chartreuse, turquoise, honeydew) and lowest (sienna, mauve, teal) naming consistency were similar for children and adults.

**Color Comprehension.** All adults performed perfectly at identifying the 6 highly nameable colors ( $M = 100\%$ ; Table S5). Adults were far less accurate in correctly selecting the 6 low nameability colors, identifying roughly half of the colors correctly on average ( $M = 49.6\%$ ; 95% CI = [44.8%, 54.5%]),  $t(89) = 20.75, p < .001$  (chance performance is 16.7%). Likewise, almost all children performed perfectly at identifying the 6 highly nameable colors ( $M = 99.7\%$ ; one child selected 4 out of 6 colors correctly). However, children were much less accurate in correctly selecting the 6 low nameability colors ( $M = 26.1\%$ ; 95% CI = [22.5%, 29.8%]), paired  $t$ -test  $t(96) = 39.62, p < .001$ , though children were still above chance among the low nameability colors overall,  $t(96) = 5.13, p < .001$ .

**Vocabulary Test.** Performance on the vocabulary test did not differ for participants assigned to the High Nameability condition vs. the Low Nameability condition, both among adults ( $t(88) = 0, p = 1$ ) and among children ( $t(95) = -0.21, p = .83$ ). On average, children ( $M = 77.8\%$ , 95% CI = [75.2%, 80.5%]) scored lower on the vocabulary test than adults ( $M = 98.0\%$ , 95% CI = [96.9%, 99.0%]),  $t(185) = 13.70, p < .001$ .

### **Relation between Category Learning and Color Word Knowledge**

**Color Comprehension and Category Learning.** Because word knowledge for high nameability colors had little to no variability, we did not fit any models testing the predictiveness of knowledge of high nameability color words. To investigate whether variability in participants' knowledge of low nameability color words predicted category learning accuracy, we conducted an exploratory analysis in which we fit a linear model, separately for adults and for children,

predicting overall category learning accuracy from low nameability color comprehension, condition (centered; high=0.5, low=-0.5), and their interaction.

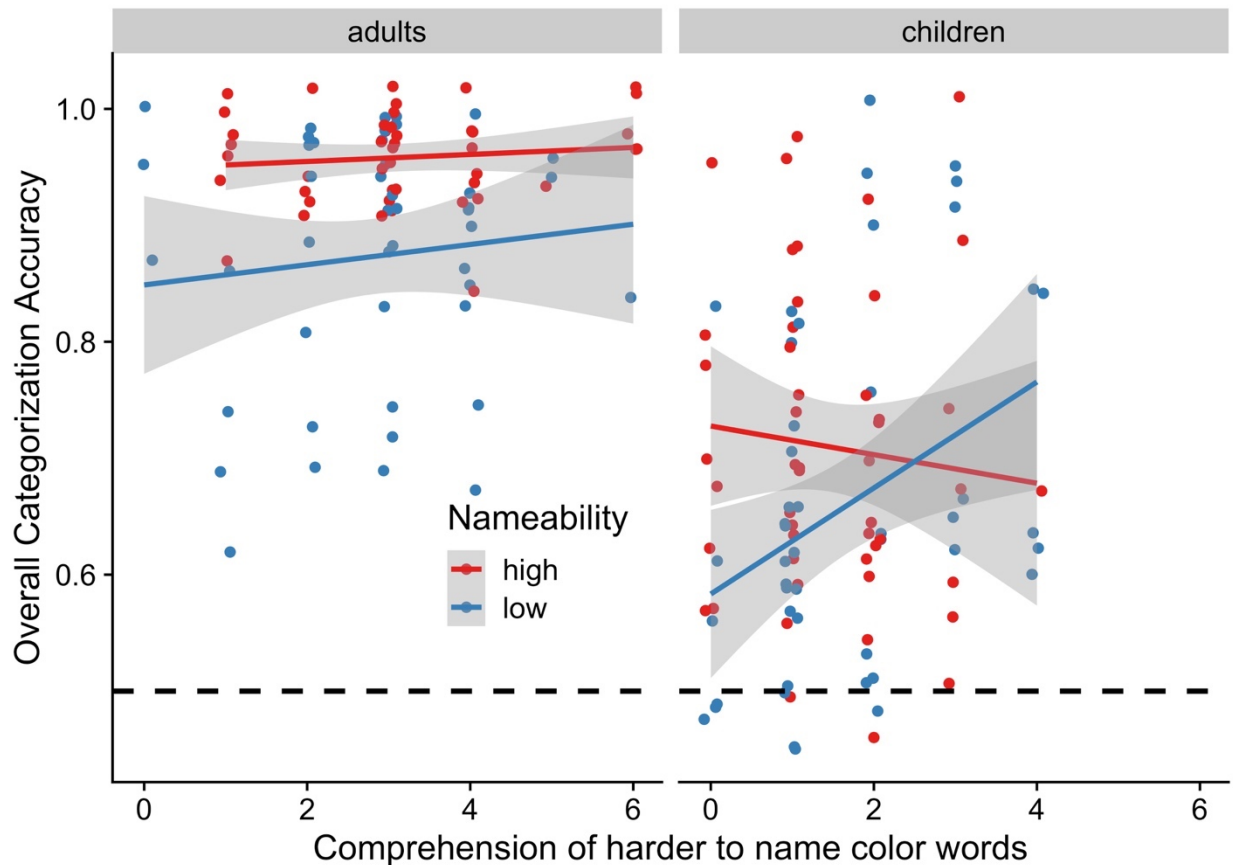


Figure 4. Relation between low nameability color word knowledge and categorization accuracy for adults and children in the High vs. Low Nameability condition.

*Adults.* There was no interaction between condition and color comprehension ( $b = -0.03$ ,  $t(86) = -0.45$ ,  $p = .65$ ) and no overall effect of low nameability color word knowledge on category learning ( $b = 0.04$ ,  $t(86) = 0.93$ ,  $p = .35$ ; Figure 4).

*Children.* We found a significant interaction between low nameability color comprehension score and condition,  $b = -0.06$ , 95% CI =  $[-0.11, -0.01]$ ,  $t(93) = -2.21$ ,  $p = .03$ , indicating that the effect of children's knowledge of low nameability color words on category learning differed across the two conditions (Figure 4). Higher word knowledge for low

nameability colors predicted better category learning performance in the low nameability condition (i.e., the condition with features composed of low nameability colors),  $b = .05$ , 95% CI = [.01, .08],  $t(93) = 2.66$ ,  $p = .009$ , while low nameability color word knowledge did not predict category learning accuracy in the high nameability condition ( $p = .54$ ). The interaction between low nameability color word knowledge and category learning accuracy remained significant when including age (as well as its interactions with other predictors) and controlling for overall vocabulary score in the model ( $t(93) = -1.99$ ,  $p = 0.0499$ ), and the interaction effect was not moderated by age ( $p = .33$ ; see Supplementary Materials S4.1. for details).

### Discussion

When faced with the task of abstracting novel category rules, both adults and children benefited from more nameable category features. In the present work, we found that one factor that helps both child and adult learners in generating and testing hypotheses about novel categories is the verbal accessibility of category dimensions. We replicated previous work demonstrating that adults learn better when categories are composed of more nameable features (Zettersten & Lupyan, 2020) and extended this result to 4-6-year-olds. Consistent with past work on the development of category learning (Mathy et al., 2015; Rabi et al., 2015), children exhibited low overall accuracy in categorizing stimuli organized by a simple one-dimensional rule, but had greater success during training when category features were more nameable. However, the degree to which nameability supported learning differed between the two age groups: adults showed larger benefits from more nameable category features (and showed substantially better accuracy in general) than children. For children, a learning boost from more nameable category features emerged only during the final training blocks of a given round. The effect of nameability was also specific to the Training Phase for both children and adults, with no

differences in performance during the Generalization Phase. Why did we find that nameable features support learning new categories during training? We explore several possible explanations in the sections that follow.

### **Why are Categories with More Nameable Features Easier to Learn?**

**Verbal Hypothesis-Testing.** One explanation for the current findings is that labels may be particularly powerful, compact representations of hypotheses about to-be-learned structures (Clark, 1998; Gentner, 2016; Lupyan & Zettersten, 2021). These compact representations (e.g., “alien A likes to eat *red* things”) make it easier for learners to formulate hypotheses about category rules (“it is about the red segment” vs. “it is about that pinkish-purplish segment”). Without these accessible, compact representations, learners may find it more difficult to formulate consistent hypotheses about category membership. In support of the notion that verbal labels play a causal role in our findings, we found that children had greater success in the low nameability condition when they had better vocabulary knowledge of the difficult-to-name color features. When children could more easily access labels for category-relevant dimensions, they were more likely to accurately learn about the novel category. Crucially, better vocabulary knowledge for low nameability colors did not predict better category learning for children in the high nameability condition; the effect of color feature vocabulary was specific to the category dimensions that a learner needed to generate hypotheses about. These findings are consistent with previous work demonstrating the causal role of verbal labels in children’s abstract reasoning (Christie & Gentner, 2014; Loewenstein & Gentner, 2005) and extend past results to the discovery of rule-based category features. They also have implications for models of category learning development by suggesting that experience-based differences in the ease of representing critical category dimensions modulates how easily children can learn rule-based categories.



Color vocabulary knowledge was not a perfect predictor of category accuracy. Although virtually all children demonstrated (unsurprisingly) robust knowledge of all of the relevant color terms for category dimensions in the high nameability condition, participants still showed only modest success at learning the novel category (~70% training accuracy, despite needing to learn how to categorize only six novel stimuli following 48 training trials with explicit feedback), especially when compared to the learning accuracy of adults on the same task (~95% training accuracy). Thus, verbal accessibility alone cannot explain the differences in category learning achieved by children (compared to the adult participants). We consider possible explanations for these differences in a later section.

**Non-Verbal Explanations.** An alternative explanation for the benefits of more nameable category features is that another experiential factor, correlated with nameability, helps participants discover the novel category structure. One possibility is that more nameable color features are also features that participants are more *familiar* with (e.g., because they have encountered them more frequently in the past), and this familiarity allows them to posit hypotheses about category membership more easily. A second possibility is that more nameable colors are also more *meaningful*, in the sense that they are more strongly associated with the learners' existing knowledge. Because we did not causally manipulate participants' experience with the underlying color features, the present study cannot rule out these alternative explanations. However, the relation between factors such as familiarity and nameability is less straightforward than it might appear: for example, attempts to quantify the frequency of colors in environmental scenes tend to find that low-saturation colors are pervasive, which are typically less nameable (Belpaeme & Bleys, 2009; see Zettersten & Lupyan, 2020 for further discussion). Moreover, to the extent to which frequency of exposure and the meaningfulness of the colors

used in the high nameability and the low nameability conditions vary, it is likely that this difference in experience is closely connected with differential naming experience. If a color appears more familiar or more meaningful to an observer, it is likely that this color is verbally encoded in the observer's language. The codability of colors shows surprising variability across languages (Majid et al., 2018), suggesting that characteristics such as familiarity, memorability, or the meaningfulness of colors are not inherent to the colors themselves – rather, they may be products of cultural experience in general, and to be structured by experience with color terms in particular (Forder & Lupyan, 2019; Goldstein, Davidoff, & Roberson, 2009; Winawer et al., 2007).

### **Why do Children show Weaker Effects than Adults?**

When comparing the performance of children to that of adults, two results stand out. First, adults show much more accurate learning than children. It is not obvious *a priori* that children should perform dramatically worse than adults on this task; the underlying category learning rule is strikingly simple in formal terms, requiring learners to only notice and remember a single feature for each category, and a number of adjustments were made to the task to increase children's motivation and provide them with unambiguous feedback. For example, trials were repeated until participants responded correctly, features were reinforced after correct responses (aliens' bodies changed colors to match the category stimulus features), and the task was given a game-like structure to increase children's engagement. Nevertheless, adults learned the categories far more rapidly. This finding builds on and extends past findings in the developmental category learning literature demonstrating that category rules that have a simple formalization may still be quite difficult for children and that rule-based category learning

undergoes significant developmental change (Huang-Pollock et al., 2011; Mathy et al., 2015; Rabi et al., 2015; Roark et al., 2023).

Second, adults show a substantially larger effect of nameability compared to children. If verbal labels are central to learning rule-based categories, one possible explanation to consider is that children simply are less familiar with color words. However, this explanation does not fully account for our results, because children were highly accurate in naming features belonging to the High Nameability condition, yet showed far worse performance in this condition than adults. Instead, it appears more likely that adults and children differ in how they approach the task. Specifically, adults may approach simple category learning tasks such as these by attempting to identify and test simple one- or two-dimensional rules. In contrast, children may approach the learning task with weaker priors about the type of solution (Gopnik et al., 2017; Lucas et al., 2014).

If children approach the learning task with weaker or more unstable priors about the kind of category rule to expect, this may explain in part why the nameability effect emerged only in the final block of a given training round among children. Children may first begin the task without necessarily crafting rule-based strategies that require representing relevant category dimensions, only seeking to generate specific rule-like hypotheses later in the task. Once children begin generating hypotheses about category rules, the nameability of category dimensions supports the ease with which they can form and test these hypotheses. Future work could test this hypothesis by scaffolding children's ability to generate rule-based hypotheses (e.g., by training children that categories will follow a simple rule based on an individual feature of each stimulus). If the children show a weaker effect of nameability mainly due to

inconsistency in seeking simple category rules, then supporting children's tendency to form rule-based hypotheses should magnify the nameability effect.

Further evidence supporting the idea that children solve the task differently than adults lies in children's and adults' behavior when tested on novel items after the training phase. The Generalization Phase was designed primarily to explore differences in categorization strategy between children and adults. The novel item was intended to distinguish participants categorizing based on a single color feature versus participants categorizing based on multiple color features. This distinction was ultimately not useful for detecting how categorization strategies differed between high and low nameability participants. The lack of a nameability effect in the Generalization Phase is likely a consequence of participants in the high nameability condition learning rule-like strategies involving both single features and multiple features (Zettersten & Lupyan, 2020), which may have masked differences in underlying strategies between the high and low nameability condition (see S5 in the Supplementary Materials for further discussion). However, inspecting response patterns for the novel items in the Generalization Phase allowed us to identify a fundamental difference in children's and adults' category learning. Adult participants showed highly consistent responses, performing at ceiling for prototype items and showing consistent sorting behavior for the novel items, with 75-90% of participants sorting the item consistently into one category or the other. Children, however, were far more variable on both item types (only 29% - 49% of participants sorted items consistently). Adults' sorting patterns suggest they were employing a consistent categorization strategy (whether based on single or multiple features), while children's more varied sorting suggests more inconsistency in the use a specific strategy across individuals and within a given testing session, in line with other findings suggesting more inconsistent responding among children in

perceptual classification tasks (Thompson, 1994). This observation is also consistent with other evidence that children can perform better than adults at some types of category learning tasks when the underlying category rule does not align with adults' expectations, because children are more likely to use novel strategies (Gopnik et al., 2017; Liquin & Gopnik, 2022; Lucas et al., 2014). If children approach the current task with fewer expectations about the types of rules that specify the category structure, this may explain why they show weaker effects of the nameability of individual features compared to adults.

### **Implications for the Development of Category Learning**

The current work has several implications for current theory and future directions in the study of category learning and how it develops. First, the current results suggest that category complexity depends importantly on past experience (and in particular past language experience) with category features. Some categories such as “red things” may be easier to learn not just because they are “inherently easy” (Feldman, 2003) or because they are grounded in a “pre-existing conceptual space” (Li & Gleitman, 2002), but because of a developmental history that makes it easier to form and test verbal hypotheses about some features than others. Even categories that have identical formal structure can vary in difficulty depending on the nameability of underlying stimulus features. A key consideration for future work will be to investigate how these findings generalize to other types of features, such as shape. Zettersten & Lupyan (2020) found similar effects of nameability for both multiple sets of color and shape features, suggesting that effects of verbalizability may apply broadly to many kinds of category features – however, it is an open question whether nameability effects among children will generalize in a similar fashion. Second, these findings predict individual variation in the ease of learning categories depending on language experience. In the current work, we find that children

with greater knowledge of words for difficult-to-name colors are more successful at learning categories composed of these color features. If language experience can make abstract, rule-based categories easier to learn, this may also explain why individual differences in vocabulary is a surprisingly strong predictor of many later educational outcomes (Bleses et al., 2016). Finally, our findings also carry intriguing implications for cross-cultural variation in category learning. Languages vary substantially in the degree to which colors – as well as other basic features such as shape – are easily verbalized (Majid et al., 2018). Our findings predict that differential experience encoding features of the world into language may systematically shift how easy it is to learn novel categories. Future work can therefore build on the current findings by investigating the emergence and developmental trajectory of cross-linguistic differences in category learning in tandem with cross-linguistic differences in vocabulary.

### **Conclusion**

Learning to categorize items according to rules is a central component of cognitive development and is important for many everyday behaviors, including navigating the environment and playing games. However, abstracting even simple rule-based categories is not trivial for children. Our results reveal one factor that influences the ease with which both adults and children form rule-based categories: the nameability of relevant category features. Words may help both adults and children learn rule-based categories, though there are likely other factors that substantially shift how adults approach the task as compared to children (Gopnik et al., 2017; Munakata et al., 2012). These factors may in turn magnify the importance of accessible abstract representations across development – the kind delivered by labels.

### References

- Anderson, E. M., Chang, Y.-J., Hespos, S., & Gentner, D. (2018). Comparison within pairs promotes analogical abstraction in three-month-olds. *Cognition*, *176*, 74–86. <https://doi.org/10.1016/j.cognition.2018.03.008>
- 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*, 442–481. <https://doi.org/10.1037/0033-295X.105.3.442>
- Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, *59*, 390–412. <https://doi.org/10.1016/j.jml.2007.12.005>
- 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*, 255–278. <https://doi.org/10.1016/j.jml.2012.11.001>
- Bates, D., Mächler, M., Bolker, B. M., & Walker, S. C. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*, 1–48. <https://doi.org/10.18637/jss.v067.i01>
- Belpaeme, T., & Bleys, J. (2009). The impact of statistical distributions of colours on colour category acquisition. *Journal of Cognitive Science*, *10*, 1–20. <https://doi.org/10.17791/jcs.2009.10.1.1>
- Best, C. A., Yim, H., & Sloutsky, V. M. (2013). The cost of selective attention in category learning: developmental differences between adults and infants. *Journal of Experimental Child Psychology*, *116*, 105–119. <https://doi.org/10.1016/j.jecp.2013.05.002>
- Bleses, D., Makransky, G., Dale, P. S., Højen, A., & Ari, B. A. (2016). Early productive vocabulary predicts academic achievement 10 years later. *Applied Psycholinguistics*, *37*, 1461–1476. <https://doi.org/10.1017/S0142716416000060>
- Carey, S. (2011). Précis of The Origin of Concepts. *Behavioral and Brain Sciences*, *34*, 113–167. <https://doi.org/10.1017/S0140525X10000919>

- Casasola, M., Bhagwat, J., & Burke, A. S. (2009). Learning to form a spatial category of tight-fit relations: How experience with a label can give a boost. *Developmental Psychology, 45*, 711–723. <https://doi.org/10.1037/a0015475>
- Christie, S., & Gentner, D. (2014). Language helps children succeed on a classic analogy task. *Cognitive Science, 38*, 383–397. <https://doi.org/10.1111/cogs.12099>
- Clark, A. (1998). Magic words: How language augments human computation. In P. Carruthers & J. Boucher (Eds.), *Language and Thought: Interdisciplinary Themes* (pp. 162–183). Cambridge University Press. <https://doi.org/10.1017/CBO9780511597909.011>
- 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*, 172–183. <https://doi.org/10.1037/a0016748>.Rules
- de Leeuw, J. R. (2015). jsPsych: A JavaScript library for creating behavioral experiments in a Web browser. *Behavior Research Methods, 47*, 1–12. <https://doi.org/10.3758/s13428-014-0458-y>
- Deng, W. (Sophia), & Sloutsky, V. M. (2016). Selective attention, diffused attention, and the development of categorization. *Cognitive Psychology, 91*, 24–62. <https://doi.org/10.1016/j.cogpsych.2016.09.002>
- Dunn, L. M., & Dunn, D. M. (2007). Peabody picture vocabulary test (4th ed.). *Bloomington: NCS Pearson*.
- Feldman, J. (2003). What is a visual object? *Trends in Cognitive Sciences, 7*, 252–256. [https://doi.org/10.1016/S1364-6613\(03\)00111-6](https://doi.org/10.1016/S1364-6613(03)00111-6)
- Forbes, S. H., & Plunkett, K. (2018). Linguistic and cultural variation in early color word learning. *Child Development, 91*, 28–42. <https://doi.org/10.1111/cdev.13164>
- Forbes, S. H., & Plunkett, K. (2019). Infants show early comprehension of basic color words. *Developmental Psychology, 55*, 240–249. <https://doi.org/10.1037/dev0000609>
- 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, 1105–1123. <https://doi.org/10.1037/xge0000560>
- Frank, M. C., Everett, D. L., Fedorenko, E., & Gibson, E. (2008). Number as a cognitive technology: Evidence from Pirahã language and cognition. *Cognition*, *108*, 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*, 650–657. <https://doi.org/10.1037/amp0000082>
- Gentner, D., & Christie, S. (2010). Mutual bootstrapping between language and analogical processing. *Language and Cognition*, *2*, 261–283. <https://doi.org/10.1515/langcog.2010.011>
- Gentner, D., Ozyürek, A., Gürçanlı, O., & Goldin-Meadow, S. (2013). Spatial language facilitates spatial cognition: Evidence from children who lack language input. *Cognition*, *127*, 318–330. <https://doi.org/10.1016/j.cognition.2013.01.003>
- Goldstein, J., Davidoff, J., & Roberson, D. (2009). Knowing color terms enhances recognition: Further evidence from English and Himba. *Journal of Experimental Child Psychology*, *102*, 219–238. <https://doi.org/10.1016/j.jecp.2008.06.002>
- Gopnik, A., O’Grady, S., Lucas, C. G., Griffiths, T. L., Wente, A., Bridgers, S., Aboody, R., Fung, H., & Dahl, R. E. (2017). Changes in cognitive flexibility and hypothesis search across human life history from childhood to adolescence to adulthood. *Proceedings of the National Academy of Sciences*, *114*, 7892–7899. <https://doi.org/10.1073/pnas.1700811114>
- Guest, S., & Laar, D. Van. (2002). The effect of name category and discriminability on the search characteristics of colour sets. *Perception*, *31*, 445–461. <https://doi.org/10.1068/p3134>
- Hochmann, J. R., Tuerk, A. S., Sanborn, S., Zhu, R., Long, R., Dempster, M., & Carey, S. (2017). Children’s representation of abstract relations in relational/array match-to-sample tasks. *Cognitive Psychology*, *99*, 17–43. <https://doi.org/10.1016/j.cogpsych.2017.11.001>
- Huang-Pollock, C. L., Maddox, W. T., & Karalunas, S. L. (2011). Development of implicit and explicit category learning. *Journal of Experimental Child Psychology*, *109*, 321–335. <https://doi.org/10.1016/j.jecp.2011.02.002>
- Jaeger, T. F. (2008). Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models. *Journal of Memory and Language*, *59*, 434–446.

<https://doi.org/10.1016/j.jml.2007.11.007>

- Kloos, H., & Sloutsky, V. M. (2008). What's behind different kinds of kinds: Effects of statistical density on learning and representation of categories. *Journal of Experimental Psychology. General*, *137*, 52–72. <https://doi.org/10.1037/0096-3445.137.1.52>
- 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*, 552–572. <https://doi.org/10.1037/a0029178>
- Li, P., & Gleitman, Li. (2002). Turning the tables: Language and spatial reasoning. *Cognition*, *83*, 265–294. [https://doi.org/10.1016/S0010-0277\(02\)00009-4](https://doi.org/10.1016/S0010-0277(02)00009-4)
- Lillard, A. S., Lerner, M. D., Hopkins, E. J., Dore, R. A., Smith, E. D., & Palmquist, C. M. (2013). The impact of pretend play on children's development: A review of the evidence. *Psychological Bulletin*, *139*, 1–34. <https://doi.org/10.1037/a0029321>
- Liquin, E., & Gopnik, A. (2022). Children are more exploratory and learn more than adults in an approach-avoid task. *Cognition*, *218*, 104940. <https://doi.org/10.1016/j.cognition.2021.104940>
- Loewenstein, J., & Gentner, D. (2005). Relational language and the development of relational mapping. *Cognitive Psychology*, *50*, 315–353. <https://doi.org/10.1016/j.cogpsych.2004.09.004>
- Long, B. L., Fan, J. E., Chai, Z., & Frank, M. C. (2019). Developmental changes in the ability to draw distinctive features of object categories. *Proceedings of the 41st Annual Conference of the Cognitive Science Society*. <https://doi.org/10.1167/19.10.59b>
- Long, B. L., Fan, J. E., & Frank, M. C. (2018). Drawings as a window into developmental changes in object representations. *Proceedings of the 40th Annual Conference of the Cognitive Science Society*.
- Lucas, C. G., Bridgers, S., Griffiths, T. L., & Gopnik, A. (2014). When children are better (or at least more open-minded) learners than adults: Developmental differences in learning the forms of causal relationships. *Cognition*, *131*, 284–299.

<https://doi.org/10.1016/j.cognition.2013.12.010>

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

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

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

Lupyan, G., & Zettersten, M. (2021). Does vocabulary help structure the mind? In M. D. Sera & M. Koenig (Eds.), *Minnesota Symposia on Child Psychology: Human Communication: Origins, Mechanisms, and Functions, Volume 40* (pp. 160–199). John Wiley & Sons. <https://doi.org/https://doi.org/10.1002/9781119684527.ch6>

Majid, A., Roberts, S. G., Cilissen, L., Emmorey, K., Nicodemus, B., O’Grady, L., Woll, B., LeLan, B., de Sousa, H., Cansler, B. L., Shayan, S., de Vos, C., Senft, G., Enfield, N. J., Razak, R. A., Fedden, S., Tufvesson, S., Dingemanse, M., Ozturk, O., ... Levinson, S. C. (2018). Differential coding of perception in the world’s languages. *Proceedings of the National Academy of Sciences*, 115, 11369–11376. <https://doi.org/10.1073/pnas.1720419115>

Mash, C. (2006). Multidimensional shape similarity in the development of visual object classification. *Journal of Experimental Child Psychology*, 95, 128–152. <https://doi.org/10.1016/j.jecp.2006.04.002>

Mathy, F., Friedman, O., Courenq, B., Laurent, L., & Millot, J.-L. (2015). Rule-based category use in preschool children. *Journal of Experimental Child Psychology*, 131, 1–18. <https://doi.org/10.1016/j.jecp.2014.10.008>

Miller, H. E., Patterson, R., & Simmering, V. R. (2016). Language supports young children’s use of spatial relations to remember locations. *Cognition*, 150, 170–180. <https://doi.org/10.1016/j.cognition.2016.02.006>

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*, 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 (Ed.), *Psychology of Learning and Motivation - Advances in Research and Theory* (Vol. 52, Issue C, pp. 117–162). Academic Press.  
[https://doi.org/10.1016/S0079-7421\(10\)52003-6](https://doi.org/10.1016/S0079-7421(10)52003-6)
- Munakata, Y., Snyder, H. R., & Chatham, C. H. (2012). Developing cognitive control: Three key transitions. *Current Directions in Psychological Science*, *21*, 71–77.  
<https://doi.org/10.1177/0963721412436807>
- Munroe, R. P. (2010). Color Survey Results. In *xkcd*.  
<https://doi.org/https://blog.xkcd.com/2010/05/03/color-survey-results/>
- Overkott, C., Souza, A. S., & Morey, C. C. (2023). The developing impact of verbal labels on visual memories in children. *Journal of Experimental Psychology: General*, *152*, 825–838.  
<https://doi.org/10.1037/xge0001305>
- 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>
- Plebanek, D. J., & Sloutsky, V. M. (2017). Costs of selective attention: When children notice what adults miss. *Psychological Science*, *28*, 723–732.  
<https://doi.org/10.1177/0956797617693005>
- R Development Core Team. (2022). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. <https://www.r-project.org/>
- Rabi, R., Miles, S. J., & Minda, J. P. (2015). Learning categories via rules and similarity: Comparing adults and children. *Journal of Experimental Child Psychology*, *131*, 149–169.  
<https://doi.org/10.1016/j.jecp.2014.10.007>
- Rabi, R., & Minda, J. P. (2014). Rule-based category learning in children: The role of age and executive functioning. *PLoS ONE*, *9*. <https://doi.org/10.1371/journal.pone.0085316>

- Roark, C. L., Lescht, E., Hampton Wray, A., & Chandrasekaran, B. (2023). Auditory and visual category learning in children and adults. *Developmental Psychology*.  
<https://doi.org/10.1037/dev0001525>
- Saji, N., Imai, M., & Asano, M. (2020). Acquisition of the meaning of the word orange requires understanding of the meanings of red, pink, and purple: Constructing a lexicon as a connected system. *Cognitive Science*, *44*, e12813. <https://doi.org/10.1111/cogs.12813>
- Schneider, R. M., Sullivan, J., Marušič, F., Žaucer, R., Biswas, P., Mišmaš, P., Plesničar, V., & Barner, D. (2020). Do children use language structure to discover the recursive rules of counting? *Cognitive Psychology*, *117*, 101263.  
<https://doi.org/10.1016/j.cogpsych.2019.101263>
- Schyns, P. G., Goldstone, R. L., & Thibaut, J. P. (1998). The development of features in object concepts. *The Behavioral and Brain Sciences*, *21*, 1–54.  
<https://doi.org/10.1017/S0140525X98000107>
- Schyns, P. G., & Rodet, L. (1997). Categorization creates functional features. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *23*, 681–696.  
<https://doi.org/10.1037/0278-7393.23.3.681>
- Shepard, R. N., Hovland, C. I., & Jenkins, H. M. (1961). Learning and memorization of classifications. *Psychological Monographs: General and Applied*, *75*, 1–42.  
<https://doi.org/10.1037/h0093825>
- Simms, N. K., & Gentner, D. (2019). Finding the middle: Spatial language and spatial reasoning. *Cognitive Development*, *50*, 177–194. <https://doi.org/10.1016/j.cogdev.2019.04.002>
- Simpson, E. H. (1949). Measurement of diversity. *Nature*, *163*, 688.  
<https://doi.org/10.1038/163688a0>
- Singley, A. T. M., & Bunge, S. A. (2014). Neurodevelopment of relational reasoning: Implications for mathematical pedagogy. *Trends in Neuroscience and Education*, *3*, 33–37.  
<https://doi.org/10.1016/j.tine.2014.03.001>
- Thompson, L. A. (1994). Dimensional strategies dominate perceptual classification. *Child Development*, *65*, 1627–1645.

- Visser, I., & Raijmakers, M. E. J. (2012). Developing representations of compound stimuli. *Frontiers in Psychology, 3*, 1–11. <https://doi.org/10.3389/fpsyg.2012.00073>
- Wagner, K., Dobkins, K., & Barner, D. (2013). Slow mapping: Color word learning as a gradual inductive process. *Cognition, 127*, 307–317. <https://doi.org/10.1016/j.cognition.2013.01.010>
- Wagner, K., Jergens, J., & Barner, D. (2018). Partial color word comprehension precedes production. *Language Learning and Development, 14*, 241–261. <https://doi.org/10.1080/15475441.2018.1445531>
- Walker, C. M., & Gopnik, A. (2014). Toddlers infer higher-order relational principles in causal learning. *Psychological Science, 25*, 161–169. <https://doi.org/10.1177/0956797613502983>
- Winawer, J., Witthoft, N., Frank, M. C., Wu, L., Wade, A. R., & Boroditsky, L. (2007). Russian blues reveal effects of language on color discrimination. *Proceedings of the National Academy of Sciences of the United States of America, 104*, 7780–7785. <https://doi.org/10.1073/pnas.0701644104>
- Yurovsky, D., Wagner, K., Barner, D., & Frank, M. C. (2015). Signatures of domain-general categorization mechanisms in color word learning. *Proceedings of the 37th Annual Conference of the Cognitive Science Society, 2775–2780*.
- Zelazo, P. D., & Carlson, S. M. (2012). Hot and cool executive function in childhood and adolescence: Development and plasticity. *Child Development Perspectives, 6*, 354–360. <https://doi.org/10.1111/j.1750-8606.2012.00246.x>
- Zettersten, M., & Lupyan, G. (2020). Finding categories through words: More nameable features improve category learning. *Cognition, 196*, 104135. <https://doi.org/10.1016/j.cognition.2019.104135>
- Zettersten, M., Suffill, E., & Lupyan, G. (2020). Nameability predicts subjective and objective measures of visual similarity. *Proceedings of the 42nd Annual Conference of the Cognitive Science Society*.

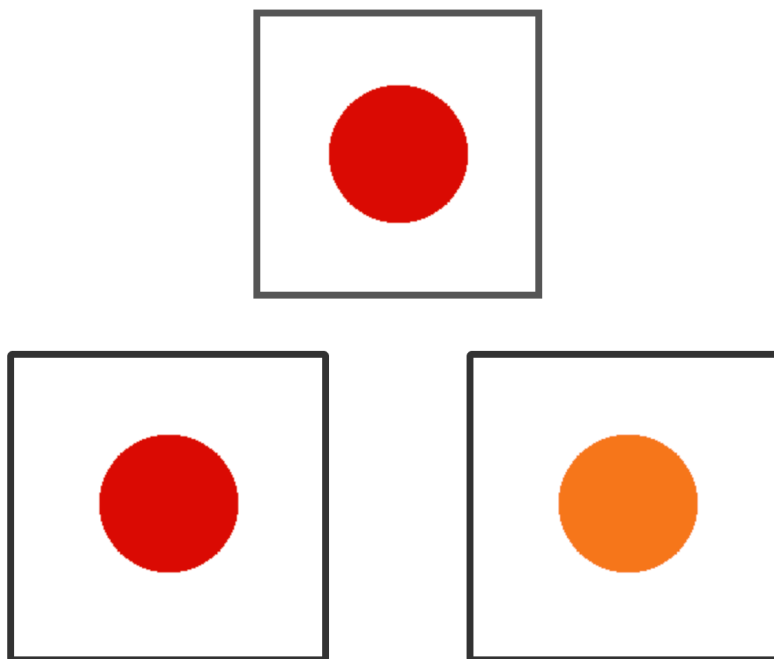
## Supplementary Materials

### S1: Additional Details on Color Feature Selection

The color features used in the current experiment were identical to those used in Zettersten & Lupyan (2020), Experiment 1B. Below, we describe the procedure originally used to identify these colors based on data from a large-scale online study of color naming (Munroe, 2010). We reproduce some of the methods descriptions from Zettersten & Lupyan (2020) in describing the selection procedure. The 12 color features (6 with high nameability; 6 with low nameability) for the 4 prototype stimuli (2 with high nameability, 2 with low nameability; each composed of three color features) were selected from among the colors with high nameability and the colors with low nameability (as quantified using Simpson's diversity index). As described in Zettersten & Lupyan (2020), the goal was to select prototype features such that the three color features of each prototype stimulus had approximately equivalent pairwise CIE-LAB distances as quantified via  $\Delta E_{2000}$  (Sharma et al., 2005). Sets of high nameability and low nameability colors were equated on CIE-LAB distances to ensure that the highly nameable color features and the low nameability color features were approximately equally easy to discriminate from one another. The resulting set of prototype images are aligned on between-color perceptual discriminability according to the following constraints: each of the three colors are clearly discriminable from the remaining two colors ( $\Delta E_{2000} > 20$ ) and the average  $\Delta E_{2000}$  values for the three color features of each prototype stimulus lie between 35 and 45. The average within-prototype feature  $\Delta E_{2000}$  discriminability was similar for high nameability colors ( $M = 39.7$ ,  $SD = 11.6$ ) and for low nameability colors ( $M = 36.5$ ,  $SD = 9.5$ ),  $t(10) = .52$ ,  $p = .61$ .

## S2. Color Norming Task

### S2.1. Trial Illustration



**Find the picture that matches the top one as fast as you can.**

*Figure S1.* Example trial in the speeded match-to-sample task



S2.2. Average Color Reaction Times

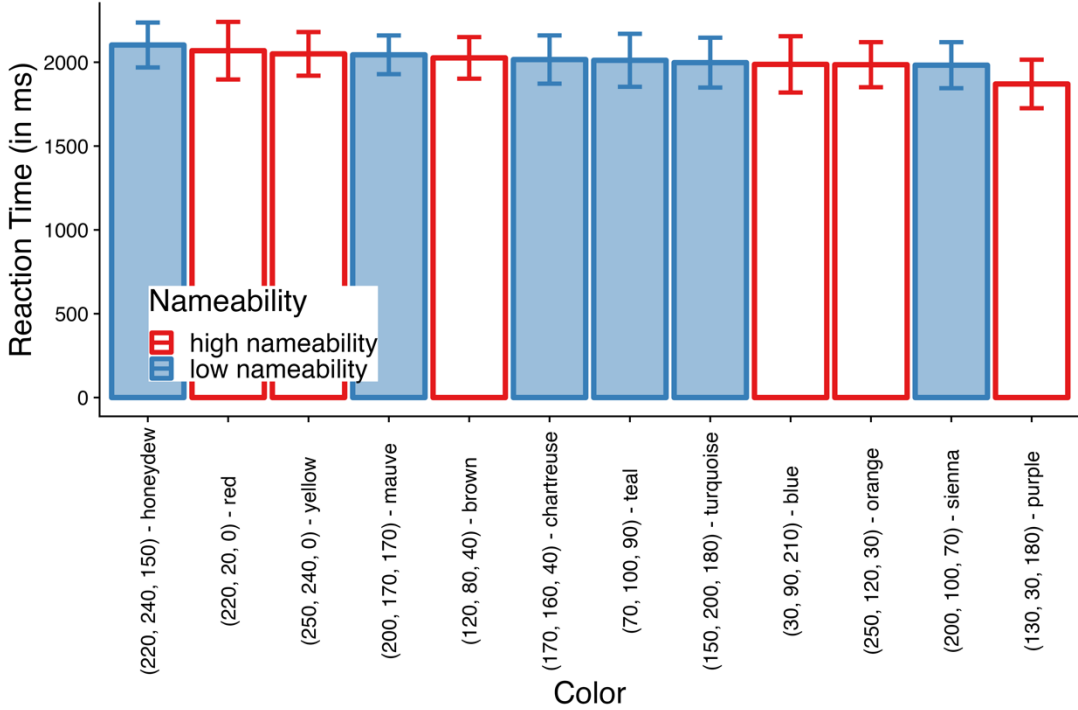


Figure S2. Average reaction times in the color norming task for children

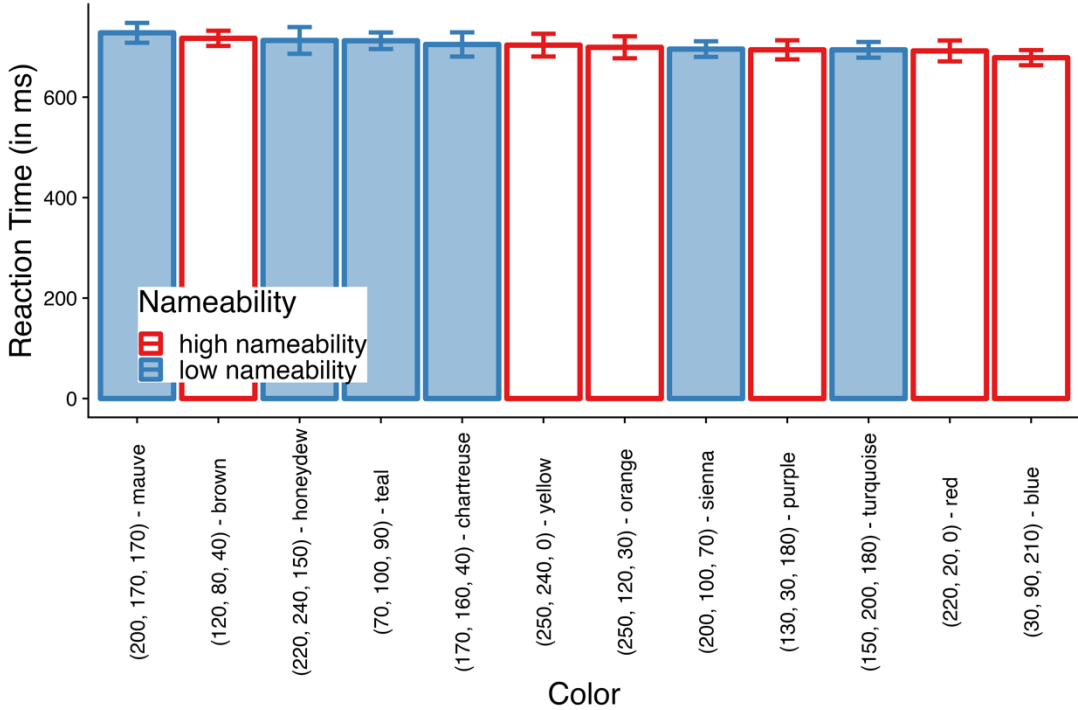
































Figure S3. Average reaction times in the color norming task for adults

### S2.3. Average Pairwise Discriminability

#### Supplementary Table 1.

*Perceptual discriminability values from the color norming task and Zettersten & Lupyan (2020)*

RGB Pair	Color Pair	Modal Names	Name-ability	$\Delta E_{2000}$	Average RT (in ms)	Average RT (in ms)	Average RT (in ms)
					Zettersten & Lupyan (2020)	Norming Task, Children	Norming Task, Adults
(30, 90, 210) - (120, 80, 40)		blue - brown	high	46	576	2047	698
(30, 90, 210) - (250, 120, 30)		blue - orange	high	56	546	1797	666
(30, 90, 210) - (130, 30, 180)		blue - purple	high	21	553	2105	705
(30, 90, 210) - (220, 20, 0)		blue - red	high	48	569	1911	672
(30, 90, 210) - (250, 240, 0)		blue - yellow	high	83	544	1981	702
(120, 80, 40) - (250, 120, 30)		brown - orange	high	31	569	1844	699
(120, 80, 40) - (130, 30, 180)		brown - purple	high	46	596	2067	706
(120, 80, 40) - (220, 20, 0)		brown - red	high	24	578	2133	694
(120, 80, 40) - (250, 240, 0)		brown - yellow	high	52	569	1945	721
(250, 120, 30) - (130, 30, 180)		orange - purple	high	58	579	2017	695
(250, 120, 30) - (220, 20, 0)		orange - red	high	21	607	2108	737
(250, 120, 30) - (250, 240, 0)		orange - yellow	high	42	579	2064	705
(130, 30, 180) - (220, 20, 0)		purple - red	high	44	597	1707	699
(130, 30, 180) - (250, 240, 0)		purple - yellow	high	93	562	1710	697
(220, 20, 0) - (250, 240, 0)		red - yellow	high	62	583	1899	664
(70, 100, 90) - (200, 100, 70)		grey green - brown	low	40	549	2160	703
(70, 100, 90) - (170, 160, 40)		grey green - mustard	low	36	583	1921	713
(70, 100, 90) - (220, 240, 150)		grey green - pale green	low	47	552	2042	675
(70, 100, 90) - (200, 170, 170)		grey green - grey	low	43	600	2165	716
(70, 100, 90) - (150, 200, 180)		grey green - green	low	34	592	1896	709
(200, 100, 70) - (170, 160, 40)		brown - mustard	low	37	538	2021	696
(200, 100, 70) - (220, 240, 150)		brown - pale green	low	50	558	1946	725
(200, 100, 70) - (200, 170, 170)		brown - grey	low	24	582	2008	733
(200, 100, 70) - (150, 200, 180)		brown - green	low	47	544	1917	680
(170, 160, 40) - (220, 240, 150)		mustard - pale green	low	21	643	1954	764
(170, 160, 40) - (200, 170, 170)		mustard - grey	low	32	588	2025	687
(170, 160, 40) - (150, 200, 180)		mustard - green	low	28	587	1916	698
(220, 240, 150) - (200, 170, 170)		pale green - grey	low	35	552	2055	706
(220, 240, 150) - (150, 200, 180)		pale green - green	low	21	589	2108	700
(200, 170, 170) - (150, 200, 180)		grey - green	low	35	584	2026	716

## S2.4. Correlations between Behavioral Discriminability Norms and CIE-LAB distances

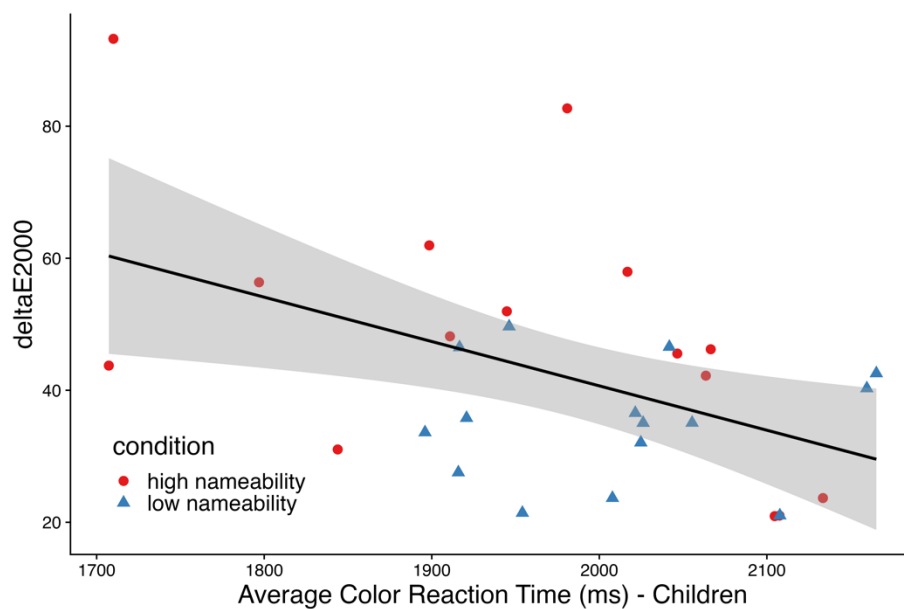


Figure S4. Correlation between pairwise average reaction times in the color norming task for children and  $\Delta E_{2000}$

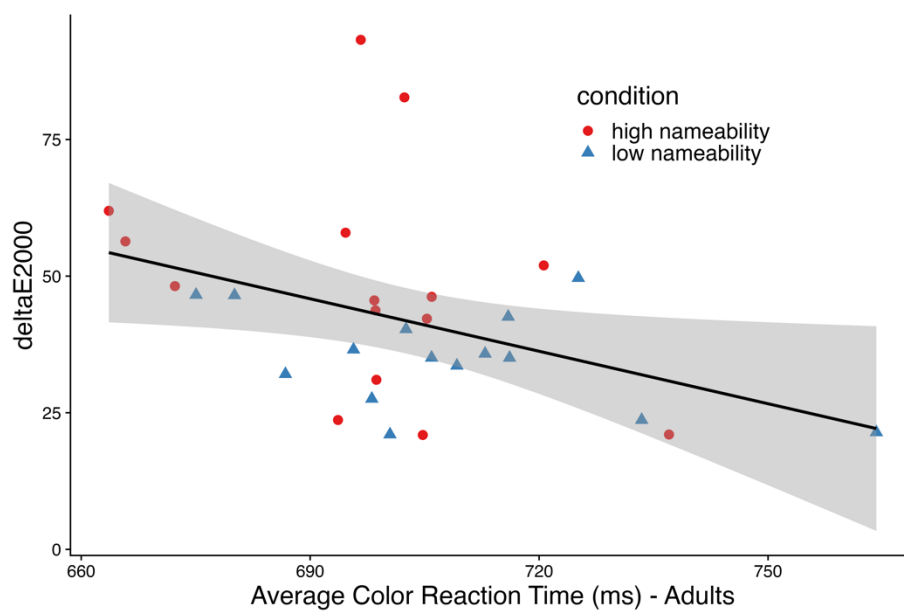
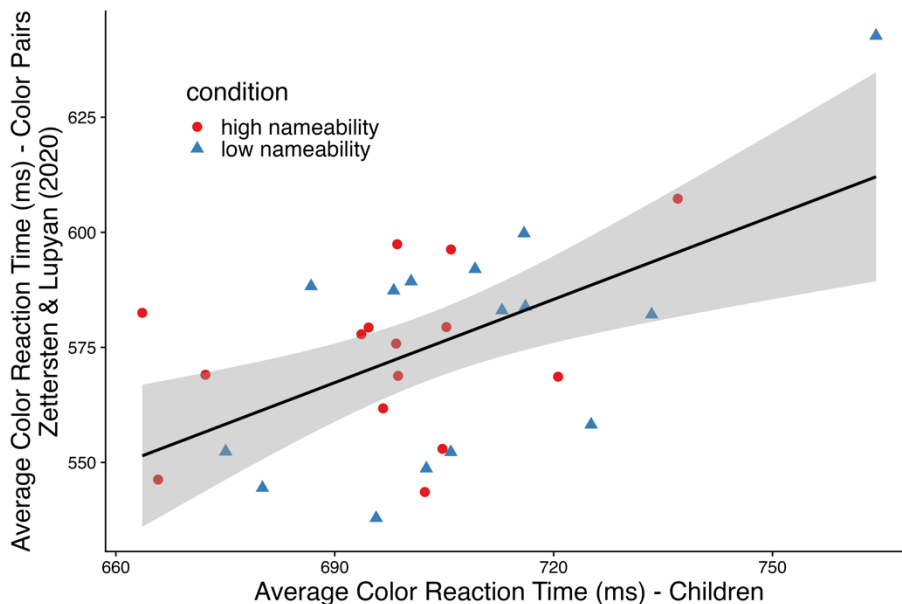


Figure S5. Correlation between pairwise average reaction times in the color norming task for adults and  $\Delta E_{2000}$



*Figure S6.* Correlation between pairwise average reaction times in the color norming task for adults and discriminability norms from Zettersten & Lupyan (2020)

### S3. Additional Modeling Details – Training Phase

#### S3.1. Comparing children and adults

To compare category learning accuracy during the Training Phase for children and adults, we fit a logistic mixed-effects model predicting trial-by-trial accuracy from Condition (centered; Low Nameability = -0.5, High Nameability = 0.5), Block Number (centered), Experiment Round (centered), Age Group (children = -0.5; adults = 0.5), and all possible interactions. We included the maximal by-subject random effects structure, including a by-subject random intercept and a by-subject random slope for Block Number, Experiment Round, and their interaction. Table S2 provides an overview of the model coefficient estimates.

**Supplementary Table 2.** *Estimates for the model comparing children and adults*

<b>Coefficient</b>	<b>Estimate</b>	<b>SE</b>	<b>z</b>	<b>p</b>
<b>Intercept</b>	2.11	0.09	23.09	<.001
<b>Condition</b>	0.93	0.17	5.38	<.001
<b>Block Number</b>	0.55	0.07	7.86	<.001
<b>Round</b>	0.79	0.12	6.42	<.001
<b>Age Group</b>	2.32	0.18	13.18	<.001
<b>Condition * Block Number</b>	0.37	0.12	3.18	.001
<b>Condition * Round</b>	0.18	0.21	0.84	.40
<b>Condition * Age Group</b>	1.28	0.34	3.73	<.001
<b>Block Number * Round</b>	-0.36	0.13	-2.65	.008
<b>Block Number * Age Group</b>	0.78	0.12	6.37	<.001
<b>Round * Age Group</b>	0.68	0.22	3.04	.002
<b>Condition * Block Number * Round</b>	-0.17	0.23	-0.74	.46
<b>Condition * Block Number * Age Group</b>	0.28	0.24	1.19	.24
<b>Condition * Round * Age Group</b>	0.22	0.43	0.52	.60
<b>Block Number * Round * Age Group</b>	-0.25	0.24	-1.06	.29
<b>Condition * Block Number * Round * Age Group</b>	-0.09	0.45	-0.19	.85

### S3.2. Interaction with child age

To investigate whether category learning accuracy during the Training Phase changed across age for children, we fit a logistic mixed-effects model predicting trial-by-trial accuracy from Condition (centered; Low Nameability = -0.5, High Nameability = 0.5), Block Number (centered), Experiment Round (centered), Age (centered), and all possible interactions. We included the maximal by-subject random effects structure, including a by-subject random

intercept and a by-subject random slope for Block Number, Experiment Round, and their interaction. Table S3 provides an overview of the model coefficient estimates.

**Supplementary Table 3.** *Estimates for the model including the interaction with child age*

<b>Coefficient</b>	<b>Estimate</b>	<b>SE</b>	<b>z</b>	<b>p</b>
<b>Intercept</b>	0.97	0.09	11.18	<.001
<b>Condition</b>	0.25	0.17	1.47	.14
<b>Block Number</b>	0.13	0.05	2.51	.01
<b>Round</b>	0.48	0.12	4.84	<.001
<b>Age</b>	0.04	0.01	3.35	<.001
<b>Condition * Block Number</b>	0.24	0.09	2.58	.01
<b>Condition * Round</b>	0.05	0.19	0.28	.78
<b>Condition * Age</b>	0.004	0.02	0.19	.85
<b>Block Number * Round</b>	-0.26	0.10	-2.69	.007
<b>Block Number * Age</b>	0.01	0.01	1.32	.19
<b>Round * Age</b>	0.02	0.01	1.65	.10
<b>Condition * Block Number * Round</b>	-0.14	0.18	-0.80	.42
<b>Condition * Block Number * Age</b>	0.003	0.01	0.20	.84
<b>Condition * Round * Age</b>	0.01	0.03	0.54	.59
<b>Block Number * Round * Age</b>	-0.01	0.01	-.90	.37
<b>Condition * Block Number * Round * Age</b>	-0.02	0.02	-0.93	.35

#### **S4. Additional Modeling Details - Relation between Category Learning and Color Word Knowledge**

##### **S4.1. Color comprehension and category learning: Interaction with child age**

We also explored whether the interaction between low nameability color comprehension and condition in predicting category learning accuracy depended on child age. To test this question,

we fit a linear model predicting children’s category learning accuracy, for low nameability color comprehension, condition (centered; high=0.5, low=-0.5), child age (centered), and all interactions between these three predictors, while also controlling for children’s vocabulary test score. Table S4 provides an overview of the model coefficients. As in the main model (not including age), there was a significant interaction between low nameability color comprehension and condition,  $b = -0.05$ , 95% CI = [-0.11, -0.01,],  $t(93) = -1.99$ ,  $p = .0499$ .

**Supplementary Table 4.** *Estimates for the three-way interaction model predicting category learning accuracy from low nameability color comprehension score, condition, and age.*

<b>Coefficient</b>	<b>Estimate</b>	<b>SE</b>	<b><i>t</i></b>	<b><i>p</i></b>
<b>Intercept</b>	0.65	0.09	7.47	<.001
<b>Condition</b>	0.14	0.05	2.72	.008
<b>Low Nameability Color Comp.</b>	0.01	0.01	1.09	.28
<b>Age</b>	0.004	0.004	1.21	.23
<b>Vocabulary Score</b>	0.01	0.11	0.13	.90
<b>Condition * Low Nameability Color Comp.</b>	-0.05	0.03	-1.99	.0499
<b>Condition * Age</b>	0.007	0.007	1.06	.29
<b>Low Nameability Color Comp. * Age</b>	0.001	0.002	0.52	.61
<b>Condition * Low Nameability Color Comp. * Age</b>	-0.003	0.003	-0.98	.33

## S5. Category Learning Accuracy – Generalization Phase

### S5.1. Generalization Accuracy

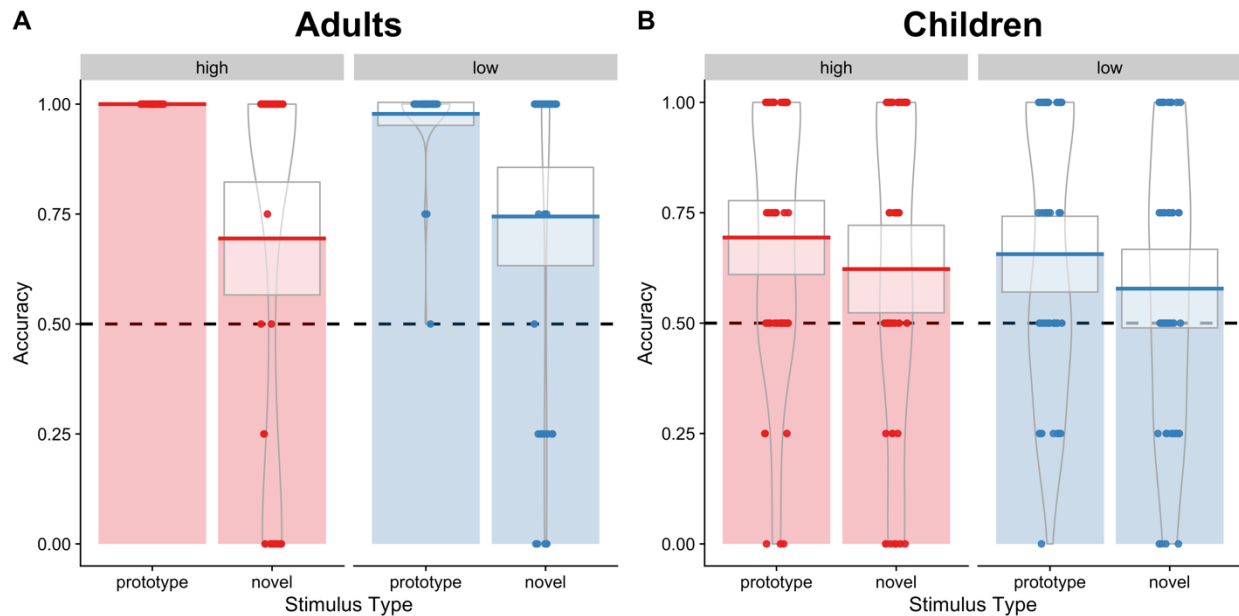


Figure S7. Generalization accuracy by condition (high vs. low) and stimulus type (prototype vs. novel) for (A) adults and (B) children. Dashed horizontal line indicates chance-level responding.

### S5.2. Sorting Consistency

We also investigated the proportion of participants who sorted the novel generalization exemplars into one category or the other, i.e. consistently according to the 100% predictive color feature, or consistently in accordance with the 75% predictive color features. Among adult participants, 88.9% of participants in the High Nameability condition (40 of 45) and 75.6% of participants in the Low Nameability condition (34 of 45) consistently sorted the novel items into one category or the other (no significant difference between conditions,  $\chi^2(1) = 1.90$ ,  $p = .17$ ). Among children, participants in the High Nameability condition (24 of 49; 49.0%) consistently sorted the novel generalization exemplars into one category or the other at a marginally higher rate than participants in the Low Nameability condition (14 of 48; 29.2%),  $\chi^2(1) = 3.21$ ,  $p = .07$ . We also compared adults and children in the consistency with which they sorted novel items into



one category or the other. Adults were far more likely to consistently assign a given 2-color difference item to a given category in both the High Nameability condition ( $\chi^2(1) = 15.41, p < .001$ ) and the Low Nameability condition ( $\chi^2(1) = 18.20, p < .001$ ).

### **S5.3. Relation between Generalization Phase Accuracy and Color Word Knowledge**

In order to investigate the effect of low nameability color word comprehension on generalization accuracy, we fit linear models, separately for adult participants and for child participants, predicting accuracy during the generalization phase from the interaction between low nameability color comprehension and condition. For both adult participants and child participants, there was no significant interaction between comprehension of difficult-to-name colors and condition in predicting accuracy in the generalization phase (adults:  $b = -0.16, t(86) = -0.81, p = .42$ ; children:  $b = -0.27, t(93) = -1.06, p = .29$ ).

### **S5.4. Why are there no effects of nameability during the Generalization Phase?**

Unlike in the training phase, there was no evidence for an effect of nameability on generalization accuracy, among adults or among children. There was a tendency in both children and adults to sort the novel items more consistently as belonging to one category or the other (i.e., to have a generalization “accuracy” of 0 or 1) in the High Nameability condition (adults: 89% of participants; children: 49% of participants) than in the Low Nameability condition (adults: 76% of participants; children: 29% of participants), but this tendency did not reach statistical significance. The most striking difference is that adults performed far more accurately on the task than children: adults performed at ceiling on the prototype stimuli by the end of the category phase, while children’s accuracy was far lower (between 65-70%). Adults also sorted novel items far more consistently into one category or the other than child participants. This









suggests that adults employed consistent strategies while performing the task, while children's strategies appear to be more mixed or inconsistent.

One possible explanation for the lack of a nameability effect is that participants in the high nameability condition and the low nameability condition may have sorted the novel items similarly, but varied in their underlying strategies. In Zettersten & Lupyan (2020), adult participants were more likely to self-report using multiple feature-based strategy in the high nameability condition than in the low nameability condition (“Circles with at least two of yellow, orange, or brown always went on the left side.”), but marginally more likely to use a holistic strategy in the low nameability condition (“Warmer colors went to the left; cooler colors went to the right”); participants reported using a single-color strategy at similar rates in both conditions by the end of the task (see Supplementary Materials S2 in Zettersten & Lupyan, 2020). The novel items were designed to be diagnostic of whether participants relied on a single feature or multiple features in making categorization decisions. The key issue is that both a holistic strategy and a category rule based on multiple color features will lead to similar categorization decisions for the novel item, which may mask differences between high and low nameability participants. In other words, generalization “accuracy” may appear similar for high nameability and low nameability participants despite underlying differences in whether participants are using more feature-based category rules or more holistic evaluations of the stimuli. Future work could further investigate this possibility by designing additional generalization items that more effectively disentangle different underlying categorization strategies among children and adults.

## S6. Overview of the Color Naming and Color Comprehension Results

**Supplementary Table 5**

*Proportion of adults' and children's correct word knowledge for the color stimuli*

RGB	Color	Target label	condition	Comp. Accuracy Children	Comp. Accuracy Adults	Child Color Naming: Simpson's Diversity	Adult Color Naming: Simpson's Diversity
(30, 90, 210)		blue	high	1.0	1.0	1.0	1.0
(250, 120, 30)		orange	high	0.99	1.0	1.0	1.0
(220, 20, 0)		red	high	1.0	1.0	0.98	1.0
(250, 240, 0)		yellow	high	0.99	1.0	0.86	1.0
(120, 80, 40)		brown	high	1.0	1.0	1.0	1.0
(130, 30, 180)		purple	high	1.0	1.0	0.94	0.98
(170,160,40)		chartreuse	low	0.15	0.21	0.32	0.20
(200, 170, 170)		mauve	low	0.46	0.74	0.15	0.10
(200, 100, 70)		sienna	low	0.12	0.41	0.18	0.18
(70, 100, 90)		teal	low	0.11	0.23	0.16	0.16
(220, 240, 150)		honeydew	low	0.32	0.72	0.42	0.30
(150, 200, 180)		turquoise	low	0.39	0.66	0.20	0.28

## S7. Supplementary Results: Relation between Category Learning and Color Word Knowledge

### S7.1. Color Naming Description Length and Category Learning

In addition to participants' color comprehension for low nameability words, we explored a further metric of individual differences in color word knowledge: the description length of participants' color naming responses. We computed the average character length of each participant's color naming responses for high and low nameability colors. We then fit a linear model (separately for adults and for children) predicting category learning accuracy from low nameability color description length, condition, and their interaction. There was no interaction between low nameability color description length and condition for adults ( $b = -0.01$ ,  $t(86) = -1.38$ ,  $p = .17$ ) or for children ( $b = -0.003$ ,  $t(93) = -0.27$ ,  $p = .79$ ). There was also no effect of low nameability color description length on category learning in the low nameability condition for adults ( $b = 0.01$ ,  $t(86) = 1.56$ ,  $p = .12$ ) or children ( $b = 0.004$ ,  $t(93) = 0.59$ ,  $p = .56$ ).

### References

- Borghans, L., Golsteyn, B. H. H., & Zölitz, U. (2015). School quality and the development of cognitive skills between age four and six. *PLoS ONE*, *10*(7), 1–20.  
<https://doi.org/10.1371/journal.pone.0129700>
- Davidson, M. C., Amso, D., Anderson, L. C., & Diamond, A. (2006). Development of cognitive control and executive functions from 4 to 13 years: Evidence from manipulations of memory, inhibition, and task switching. *Neuropsychologia*, *44*(11), 2037–2078.  
<https://doi.org/10.1016/j.neuropsychologia.2006.02.006>
- Munakata, Y., Herd, S. A., Chatham, C. H., Depue, B. E., Banich, M. T., & O'Reilly, R. C. (2011). A unified framework for inhibitory control. *Trends in Cognitive Sciences*, *15*(10),

453–459. <https://doi.org/10.1016/j.tics.2011.07.011>

Munakata, Y., Snyder, H. R., & Chatham, C. H. (2012). Developing cognitive control: Three key transitions. *Current Directions in Psychological Science*, *21*(2), 71–77.

<https://doi.org/10.1177/0963721412436807>

Munroe, R. P. (2010). Color Survey Results. In *xkcd*.

<https://doi.org/https://blog.xkcd.com/2010/05/03/color-survey-results/>

Ritchie, S. J., & Tucker-Drob, E. M. (2018). How much does education improve intelligence? A meta-analysis. *Psychological Science*, *29*(8), 1358–1369.

<https://doi.org/10.1177/0956797618774253>

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>

Zettersten, M., & Lupyan, G. (2020). Finding categories through words: More nameable features improve category learning. *Cognition*, *196*, 104135.

<https://doi.org/10.1016/j.cognition.2019.104135>